第 1 章　序論

1.1 本研究の背景

日本における機械翻訳の研究は 1955 年に九州大学の栗原らにより開始された[1,2]。その後，京都大学[3,4]を始め多くの機関で研究されるようになり，1980 年代の半ばからは製品も発売されるようになった。初期の機械翻訳は，人間により記述されたルールと辞書を用いて翻訳を行なうルール型の方式により実現された。しかし，言語表現には例外的現象が頻発するとともに，日本語と英語間の翻訳では双方の言語特性が大きく異なることから，実用レベルの翻訳品質を達成するため，ルールの精緻化を志向した様々な提案が行なわれた[5]。しかしながら，比較的小規模の語彙数に対しては効果があっても，語彙数の増大に伴って副作用が生じることも少なくなく，副作用を回避しながら大規模なルール集を完成させることは容易ではないことから，次第に行き詰まりを見せるようになった。

一方，1984 年に長尾は対訳用例を利用して翻訳を行なう用例型の方式を提唱した[6]が，ルール型の方式が全盛の時期にはほとんど顧みられなかった。ルール型の行き詰まりの中で，1989 年に佐藤らが具体的なシステムを提案した[7]ことがきっかけとなって用例型の方式に注目が集まり始め，今度は用例型翻訳一辺倒の時期が現れた。用例型の方法は，対訳用例を追加することにより翻訳品質を向上させることができるというのが利点とされている。しかし，対訳用例の必要量が明らかではなく翻訳品質の目標が定めにくいこと，単に対訳データを追加しただけでは翻訳品質の低下を招くことも多いため目的に応じた対訳データの収集が必要であること，また，その収集自体がルール型の辞書構築と同様に，あるいは，それ以上にコストがかかること，収集したデータに均質な言語情報付与を行なうのは容易ではないことなどから，実用レベルには至っていない。

統計型の方式としては，1990 年に英仏翻訳を対象とした基本モデルが提案されたが[8]，言語特性が類似する言語間の翻訳への適用に限定され，言語性質の異なる日英翻訳への適用は困難であると考えられていた。音声認識におけるパラダイムシフト，すなわち，ルール型の手法から隠れマルコフ型言語モデルを用いる手法への転換が機械翻訳にも波及し，日英翻訳にも適用されるようになってきた[9]。対訳データの収集には課題が残るものの，利用可能な対訳データ量が増えつつあることと，用例型のようにデータへの詳細な言語情報付与が必要でないこと，翻訳を行なうための言語モデルの再構築が容易であることから，盛んに行なわれるようになっている。ところで，日本語の 1 文字あたりの情報量は約 4.3bit であると推定されている[10]。このことから，記述文としては比較的短い 20 文字程度の文を対象として考える場合，そのバリエーションは 4 億文程度と推計される。新聞記事のように平均 40 文字の文では統計的に有意な効果を発揮するだけのデータ量を確保するのは極めて困難であると考えられる。したがって，現在の状況では限定的なタスクへの適用に限られると考えられる。また，データ量が確保されたとしても，現在の計算機速度では言語モデルの再構築は現実的ではない。

以上の方式は実現形態は大きく異なるが，翻訳のための言語知識の構築方法の違いであり，また，その利用方法の違いであると考えられる。すなわち，規則型は，主として経験的には整理されてきた伝統文法や辞書を電子化し，実験により不足する文法や辞書を補完することにより，言語知識として明示的に構築して利用する方式である。用例型は，言語表現を，文法や語彙の複合体として捉え，そのバリエーションを可能な限り収集しておくことにより，直接利用しようとする方法である。統計型は，文法や語彙に相当する情報を統計処理により言語モデルとして対訳データから抽出し利用する方法である。上述した表現のバリエーションの推計値から考えると，利用可能なデータ量は極めて少ないことが明らかであるから，すべてを言語データに頼って言語知識を得るのではなく，人間の言語知識を的確に投入することが，実用的な翻訳システムを実現する上で有効であることが予想される。

1.2 本研究の目的

計算言語学は理論的モデルで言語を表現し，統一的に処理することを追求してきた。しかし，自然言語は人間社会において自然発生的に成長してきた慣習であり，それぞれの言語が成長してきた社会背景を反映している。人間の社会にはさまざまな歪みや矛盾が存在し，それを見る見方や捉え方は様々である。このため，言語ごとに表現の枠組みには違いがあるのは当然であり，それを統一的手法で処理することは困難である。すべ ての実用科学がそうであったように，言語においても，汎用的な理論を考える前に，個別の言語の科学を打ち立てることが望まれる。それには，人間による言語活動を検討するところから始める必要がある。

翻訳は人間による高度な言語活動であり，その実行過程においては，「英借文」といわれる対訳そのものを流用したり表現形式をまねる翻訳のほか，表現の分析，理解，言い換えを行なってから表現を組み立てる翻訳が臨機に選択されるとともに，多種多様な言語知識が使用されると考えられる。人間の場合，個人差はあるにせよ，多くの対訳用例に一定の抽象化を行なったうえで記憶していると考えられる。また，著者が翻訳者や通訳者にインタビューした限りでは，内容が理解できない場合にもそれなりに翻訳することがある程度は可能だということであった。

計算機の場合，大量の対訳用例をそのまま蓄積することは容易であるが，自動的に抽象化することは困難である。大量の対訳用例を収集することは容易ではないが，人間に一例を提示するとそれが手がかりとなって用例を増すことができる。また，理解のような高度な言語処理をモデル化することは困難であるが，その必要性が比較的低いこと，用例が文法や辞書が複合化されたものと捉えれば，対訳用例の蓄積により代替できると考えられる。

これらのことから，可能な限り大量の対訳用例を収集すること，それらを人間が注意深く抽象化することにより言語知識として記述し利用することを考える。また，抽象化に当たっては，人間の言語活動に可能な限り配慮することとする[11]。

日英翻訳の場合，日本語と英語の言語的な性質が大きく異なることから，結合価の考え方[12]に基づいて，述語を中心として複数の格要素が組み合わさったまとまりを表現単位として翻訳する方法を考える。この方法を実現するうえで表現単位の日英対訳の記述方法が問題となるが，述語の

異なりは 1 万語程度であるのに対し，格要素の中心となる名詞の異なりは数十万語となることから，述語は表記そのものを用い，名詞は原則として意味属性を用いて抽象化することにより，対訳用例としての側面を残しながら被覆率の向上を狙う。

日英翻訳の表現単位を抽出するには，日本語の形態素解析，構文解析が必要である。形態素解析は宮崎により高精度の処理が実現されている[13]。構文解析は，日本語は語順が比較的自由であることから係り受け解析が適しているとされているが，長文の解析精度の低さが問題であった。並列関係に伴う長文解析の精度向上は黒橋らにより解決されている。そこで，本論文では，従属節の階層性[14]に着目することにより，節の関係の解析精度の向上を検討する。

次に，日英の言語的な性質の異なりについて考える。日本語は用言中心の表現が多用されるのに対し，英語は名詞中心の表現が自然である。例えば，「彼は歩いて学校へ行った。」という日本語に対応する最も自然な英語は“He walked to school.”である。日本語の 2 動詞が英語の 1 動詞に対応するため，英語を意識して「彼は学校へ歩く。」のように書き改めることにより，目的の英語表現を得る前編集がしばしば行なわれる。しかし，この作業は一般の利用者には大きな負担であるため，その自動化について検討する。

以上により，対訳用例を分割したうえで抽象化して利用することから被覆率の高い日英翻訳の実現が期待されるが，その反面，分割や抽象化により予定外の適用が発生し翻訳精度が低下する恐れが生じている。翻訳メモリのように対訳用例をそのまま利用する方法を志向した方式検討の一環として，本論文ではテンプレート翻訳と用例型の翻訳を取り上げ，実現可能性の観点から，意味的に対応する日英の文が対応付けられている対訳データを対象として検討する。すなわち，従来の用例型の翻訳のように，文対文が一意に対応し，さらに，単語や句の対応関係，文法情報や語彙情報の付与を前提とした対訳データの存在は前提としない。

テンプレート翻訳は，対訳用例の一部を変数化し，変数に代入可能な単語や句を条件指定した翻訳テンプレートを利用して翻訳する方法である。テンプレート翻訳は被覆率は高くないが，実用レベルの翻訳品質が期待できる。翻訳テンプレートの作成は人手によって行なわれてきたが，本論文では n-gram 統計の手法[15]を用いて自動抽出を行なうための方法について検討する。このテンプレート翻訳は変数箇所が固定されているが，それを動的に決定して利用することができればさらに被覆率を向上させることが期待される。日英の文が緩い対応付けである場合にでも適用可能な用例型の方法について，今後の課題として検討する。

1.3 本論文の構成

本論文の構成は以下のとおりである。

第 2 章では，日本語の認識構造を考慮することにより，従属節の関係をルールとして記述し，それを利用することにより長文の係り受け解析精度を向上させる方法を提案する。

第 3 章では，日英の言語変換精度を向上させるために行なわれる日本語の表現を英語的な表現に書き改める前編集をルール化することにより自動化する方法を提案する。

第 4 章では，述語を中心とする表現単位を結合価パターンとして日英を対にして辞書記述し，それを利用することにより，一般的な表現から慣用的な表現まで同じ処理の枠組みで日英変換可能とする方法を検討する。この辞書に基づいて，日本語語彙大系の構文体系は作成された。

第 5 章では，対訳データから翻訳テンプレートを発見する手がかりを与える n-gram 統計分析手法を提案する。

最後に，第 6 章において本研究を総括し，結論を述べる。また，今後の課題として，日英の文が緩い対応付けである場合にでも適用可能な用例利用型翻訳の構想について述べる。

第 2 章　日本語の階層的認識構造と係り受け解析

2.1 緒言

自然言語の構文解析の方法として，従来から多くの研究が行われ，様々な方法が提案されている。それらの中で，比較的自由な語順を取る日本語の構文解析としては，省略などに強い係り受け解析の方法が適していると考えられる[16]が，長い文になると，この方法も必ずしも良い成績が得られているとは言えない。

長文に対する係り受け解析で失敗する原因は，おおよそ，並列を伴う名詞句および名詞句間の解析の曖昧さと，述語間の関係の認定の曖昧さにあると考えられる。このうち，並列名詞句の扱いについては，文節列の類似性に着目した方法が提案された[17,18,19]ほか，これを省略を含む部分的並列関係に拡張し，省略を補って並列構造を抽出する方法[23]が提案され，解析精度が向上している。多少異なる角度からは，呼応関係等の文構造を決定する要因に着目して，係り受け関係解析を局所化する方法[24]も提案されている。また，格要素と述語間の関係の曖昧さを絞り込む方法としては，意味解析段階で動詞の結合価を使用する方法等[25,26,27]があり，かなりの効果が得られている。これらに対し，述語間の係り受けに対しては効果的な方法が知られていなかった。

最近の研究結果からも分かるように，従来の日本語の構文解析では，日本語の表層的特徴を，まだ，十分には使いきってはいない[16]と考えられる。表現内容と表現構造の関係が明らかになれば，表層上の特徴に着目するだけでも，今まで以上に精度の良い構文解析が実現できると期待される。特に，述語間の関係は，文全体の構造を決める重要な関係であるが，言語過程説[1]によれば，このような文の構造には，書き手が対象をとらえて，表現していく階層的な過程が反映していると考えられる[11]。

そこで，本章では，日本語の階層的な表現過程に着目して提案された，南の 3 段階の従属節分類[14]を，さらに，従属節の意味と形式に着目して，係り述節，受け述節ともに，基本分類 13 種，細分類 4 種に分類し，それらの相互関係を整理することにより，述語間の係り受け関係を決定する方法を提案する。また，新聞記事 972 文を対象とした係り受け解析において，係り先の曖昧な述語の係り先を無作為に決定する方法，直近の述語にかける方法，従来の ALT-J/E の方法[28]，本章の方法の解析精度を比較する。

1 時枝誠記が提案した言語思想で，日本の 4 大文法の一つと言われる時枝文法の思想的背景となっている。三浦つとむが哲学的立場から考察を加え，意味論を修正した。

2.2 日本語の階層的認識構造

(1) 日本語の 4 段階認識構造

日本語の文構造については，古くから多くの研究が行われており，書き手の対象に対する認識とその表現過程が，日本文の階層的な構造に反映していることが指摘されている。

山田は，日本語文では，観念的内容は先頭に，陳述に関する部分は後方に配列された階層性を持つことを指摘した[29]。時枝は，客体に対して概念化された書き手の認識の表現（客体的表現）と主体の概念化されない判断，感情の表現（主体的表現）が入れ子構造を形成することを指摘した[30]。その後，渡辺，芳賀，服部等によって，述語の構成に関する詳細な研究が行われた[31,32,33]。これらの研究成果に基づき，林は，描叙の段階，判断の段階，表出の段階，伝達の段階の 4 段階からなる入れ子型の階層構造[34]を提案した。これは，入れ子の内側ほど客観性の高い認識が表現され，入れ子の外側ほど主観性の強い内容が表現されることを 4 段階の階層構造として整理したものである。

(2) 従属節の種類と性質

南は，従属節を考える立場から，林の 4 段階の階層構造を描叙，判断，提出，表出の 4 段階に変更した[35]。そして，日本語の動詞，助動詞を中心に構成される述語は複雑な構造が持てること，その部分に表現の段階を示す要素の多くが含まれていることに着目して，日本語の節を入れ子の各段階に対応させて，4 種類に分類した。そのうえで，南は，表出段階の表現（呼びかけ，働きかけなど）は，主節には現れても，従属節には現れないことに着目して，従属節を次の 3 種類（3 段階の階層）に分類した[36]。

A 類：「～ながら」等。ほぼ林の描叙段階に相当。
B 類：「～たら，～と，～なら，～ので，～のに，～ば，～て（従属的用法）」等。ほぼ林の判断段階に相当。
C 類：「～が（順接），～から，～けれど，～し，～て（独立した用法）」等。ほぼ林の表出段階に相当。

(3) 従属節間の依存関係

南は，さらに，上記の 3 種類の従属節間に，次の強い傾向があることを明らかにした。

A 類：他の A 類，B 類，C 類の一部となることができる。
B 類：他の B 類，C 類の一部となれるが，A 類の一部とはなれない。
C 類：他の C 類の一部とはなれるが，A 類，B 類の一部とはなれない。

2.3 日本語従属節の階層的分類と順序関係

2.3.1 問題となる日本語の従属節の種類

言語処理の立場から，節をその形態に着目して分類すると，主節のほか，連用節，連体節，引用節の 3 種の従属節に分けられる。このうち主節は，他の節の係り先にはなるが係り元にはなり得ない。また，連体節は述語の活用形などの文法的性質によって係り先が明確に決まる。引用節は，引用の助詞等を伴うことが多く，係られる側の動詞のタイプが限定されるなど，形態的にその係り先がほぼ明確である。したがって，述語間の係り受け解析で問題となるのは，係り元が連用節である場合，すなわち，連用節から連用節へ，連用節から連体節へ，連用節から引用節への 3 つの係り受け関係である。

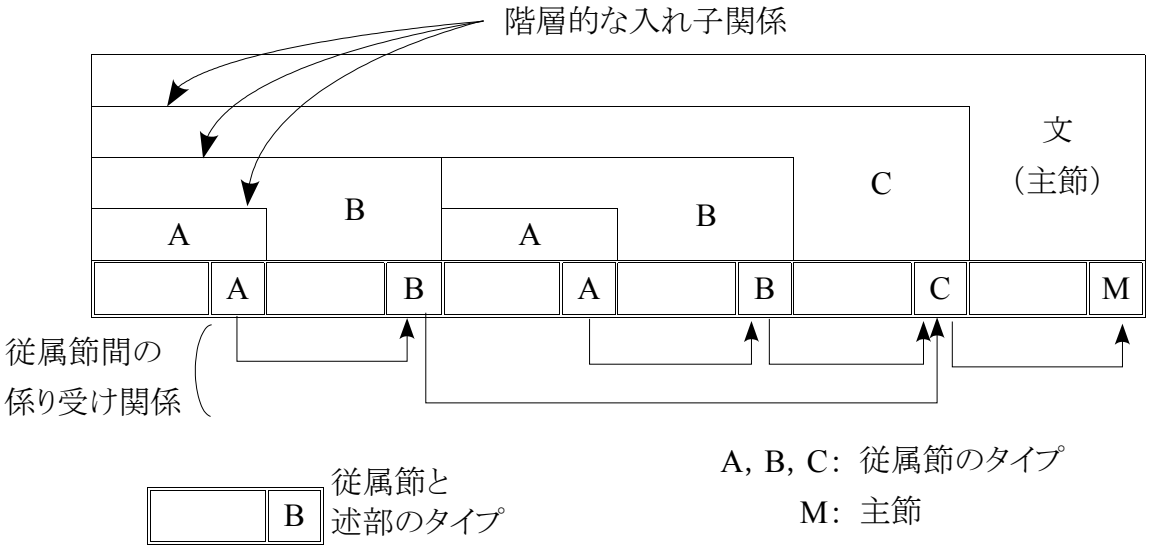


図 **2.1** 日本語述節間の階層的入れ子構造と係り受け関係

2.3.2 従属節の基本分類

(1) 従属節の基本的階層関係

本節では，まず，出現頻度が高く，係り受けが曖昧となりやすい連用節同士の係り受け関係について考える。この問題に対して，前節で述べた南の従属節 3 分類の応用を考える。2.2 節(3)の結果を従属節の包含関係から見ると，A＜B＜C の関係が成り立つ。ただし，記号“＜”は左辺が右辺に含まれる（左辺は右辺より優先度が低い）ことを意味する。また，ある節 X が他の節 Y の一部となれるということは，係り受け関係から見ると，X は Y に係ることができるということである。逆に，ある節 X が他の節 Y の一部となれないということは，X は Y に係れない，すなわち，X は Y を飛

び越えて，より後方の節に係るということである。したがって，図 **2.1** に示すように，包含関係の内側にある述語は，外側の述語に係れるが，外側にある述語は，内側にある述語には係れないことになる。

(2) 従属節再分類

上記の関係は述語間の係り受け決定に利用できると考えられるが，南の分類では，接続が順接か逆接かなど，意味的な判断が必要であるため，構文解析段階でそれを判断し，分類するのは困難である。そこで，南の分類の趣旨を生かしながら，語尾表現をより長単位で分類すること，意味的判断の困難な表記はデフォルトの解釈で分類することなどにより，従属節を次のとおり再分類する。

A 類:「同時」の表現。

B 類:「原因」,「中止」の表現。

C 類:「独立」の表現。

この分類では，表層的に明らかに A 類または C 類と判定できるもの以外は，B 類に分類した。したがって，南の分類に比べると，B 類の範囲が広くなっているが，分類相互の包含関係の傾向は保存されていると期待される。この基準で，日経産業新聞の 300 記事のリード（要約または第 1 段落，計 972 文）に現れた従属節を分類した例を表 **2.1** に示す。

表 **2.1** 新聞記事標本に現れた従属節（述部）の分類

分類	新聞記事 972 文に現れた述部［(*n*):*n* は出現回数］
A 類	～とともに(7)，～ながら(2)，～と同時に(2)，～ことに加えて(1)，～つつ(1)，～ことを含め(1)，～のをはじめ(1) 合計 7 種類，延べ度数 15 回
B 類	～(連用形単独)(159)，～て(含，「連用形+で」)(148)，～(サ変動詞の語尾省略)(94)，～で(「名詞+で」)(47)，～ため(28)，～ており(25)，～(体言止め「名詞+読点」)(20)，～ほか(16)，～ば(9)，～もので(9)，～ても(9)，～ことで(8)，～ので(4)，～と(4)，～ず(4)，～たり(3)，～ために(3)，～うえで(3)，～後(2)，～のに対応(2)，～のに続き(2)，～ためで(2)，～そうで(2)，～上で(1)，～時や(1)，～時に(1)，～際に(1)，～際(1)，～結果(1)，～以上(1)，～よう(1)，～ものの(1)，～のを手始めに(1)，～のを機に(1)，～のに応じて(1)，～ところ(1)，～ておいて(1)，～だけに(1)，～だけでなく(1)，～たら(1)，～たびに(1)，～ずに(1)，～こともあって(1)，～うえ(1)，～のに対し(1)，～なら(1) 合計 46 種類，延べ度数 626 回
C 類	～が(42) 合計 1 種類，延べ度数 42 回

～:用言（動詞，形容詞）の部分を表す

(3) 読点の有無による分類

さて，述語間の係り受けに曖昧さが生じるのは，2 つ以上の従属節がある場合である。そこで，前述の新聞記事からそのような文を抽出して，各従属節の出現頻度を調べたところ，A 類 5%，B 類 86%あまり，C 類 9%あまりで，圧倒的に B 類が多いことが分かった[2]。

そこで，書き手の習性として，解釈しにくい曖昧さのある文では，書き手自身が読点“、”を入れる傾向があること，特に，従属節を含む長文ではその傾向が強いこと，に着目して，従属節 A，B，C の分類に，さらに，読点の有無を加えた 6 種類の分類を考える。書き手が読点を付与した従属節は，それだけ，遠くに係る可能性が強い，すなわち，独立性が強いと考えられるから，それぞれの従属節の包含関係は，A＜「A+読点」＜B＜「B+読点」＜C＜「C+読点」となることが予想される。

新聞記事の例では，B 類の従属節のうち，読点を持つものと持たないものとが，ほぼ同数見られる。したがって，読点を考慮したことにより，B 類同士の係り受けの約半数は，曖昧なく決定できると期待される。

2.3.3 従属節の派生的な分類

(1) 連用節の中止性

上記の方法では，B 類同士，「B+読点」類同士の係り受け関係は決定できない。そこで，このタイプの述語について，より詳細に分類することが必要である。

そこで，B 類の述語表現を，表現の意味的な流れの中止性の強さに着目して分類することを考える。この観点から，B 類の述部を分類すると，以下のようになる。

中止性の弱いもの
　～(用言連用形)，～て，～(サ変動詞語尾省略)，～ため，～ほか，～ば，～ても
中止性の強いもの
　(名詞)で，～ており，～もので，～ことで

前者のタイプは，行為の継続，並行，順序等の意味に使用される表現であり，中止性が弱い傾向を持つのに対して，後者のタイプは，前置き的な内容を表すなど，主題の変わり目に使用される表現で，中止性が強いと考えられる[3]。

中止性の強い述語ほど遠くに係る傾向があると推定されるから，次のヒューリスティックスを導入する。

① 連用節述語のうち，中止性の強い B 類の述語は，他の連用節述語を飛び越える。逆に，他の述語は，次に現れた連用節述語に係る。

2 表 **2.1** の集計では，B 類が 92%であるが，これは係り先が曖昧でない従属節も含んでいる。係り先の問題となる述部を対象に集計すると，B 類は 86%となった。

3 この分類で，例として示した述部は，表 **2.1** に抽出された B 類述語のうち，使用傾度の高いものである。出現頻度の高い述部が，ほぼ均等に分類されていることにより，それぞれのタイプはあまり大きな偏りなく出現することが予想され，分類の効果が期待される。

② 連用形が単独で述語となっている場合は，別の単独で述語となっている連用形を飛び越えないで，それに係る。

(2) 述語の状態性と動作性

連用節のうち，B 類同士，「B+読点」類同士の係り受けを決めるため，次に，述語の状態性もしくは動作性に着目する。述語は，その動作性から見ると，動作性の強い順（状態性の弱い順）に，他動詞性，自動詞性，形容詞性，名詞性の 4 種の述語に分類することができる。ただし，使役系の表現は他動詞性，受身系の表現は自動詞性とする。

ここで，言語の表現過程を考察すると，読者に分かりやすくするため，書き手は主題や動作主体を統一的にとらえて表現する傾向がある。このため，動作性の述語と状態性の述語が同一レベルで表現されることは少ないと考えられる。また，状態性の強い述語は，動作性の強い述語に包み込まれる傾向を持つ。これらの傾向に着目して，次のヒューリスティックスを設ける。

○動作性の強い述語は，動作性の弱い述語を飛び越し，動作性の弱い述語は，動作性の強い述語に係る。

ところで，連用節の中には，用言が格助詞相当語の一部として使用される「～と比べ（～より）高い値段」のようなものがある。このタイプの連用節は，他の述語の格要素として使用されたものである。そのため，述語間の係り受けとしてではなく，用言と体言の格係り関係の一部として扱う必要がある。

2.3.4 引用節と連体節の扱い

3.2 節，3.3 節において，3.1 節で示した述語間係り受けの 3 種類の問題のうち，連用述語間の係り受けについて検討した。ここでは，残された「連用節から引用節へ」，「連用節から連体節へ」の 2 種類の係り受けについて検討する。

まず，連用節と引用節の関係を見ると，「～すると（発表する）」などの引用節は独立性が高いため，連用節が引用節を飛び越えて他の節に係ることは考えにくい。しかし，「～するよう（依頼する）」など，様態化し独立性が弱められた引用相当節の場合は，飛び越えられる可能性もある。そこで，次のように処理することとする。

① 引用節述語は，連用節述語の「C+読点」に準ずる。

② 引用相当節述語は，「B+読点」に準ずる。

次に，連用節と連体節の関係を見る。連体節を形式名詞「もの，こと，の」を伴うもの（～したものが，～することを，等）とその他の名詞性のものに分けて考えると，形式名詞を伴う連体節は，対象をとらえ直す[37]ために形式名詞が使用されたとも考えられるので，連用節がこれを越えて他の述語に係ることは考えにくい。一方，普通の名詞性の連体節の場合は，必ずしもそれが係り先になるとはいえない。

以上から，連体節述語の扱いは，次のとおりとする。

① 形式名詞に係る連体節述語は，連用節の「B+読点」に準ずる。

② 通常の連体節述語は，B に準ずる。

2.4 従属節依存関係の決定規則

2.4.1 述語種別判定

前節の結果をまとめると，表現の形態的特徴，表出過程および意味的特徴から見た日本語の従属節は，図 **2.2** のとおり分類される。

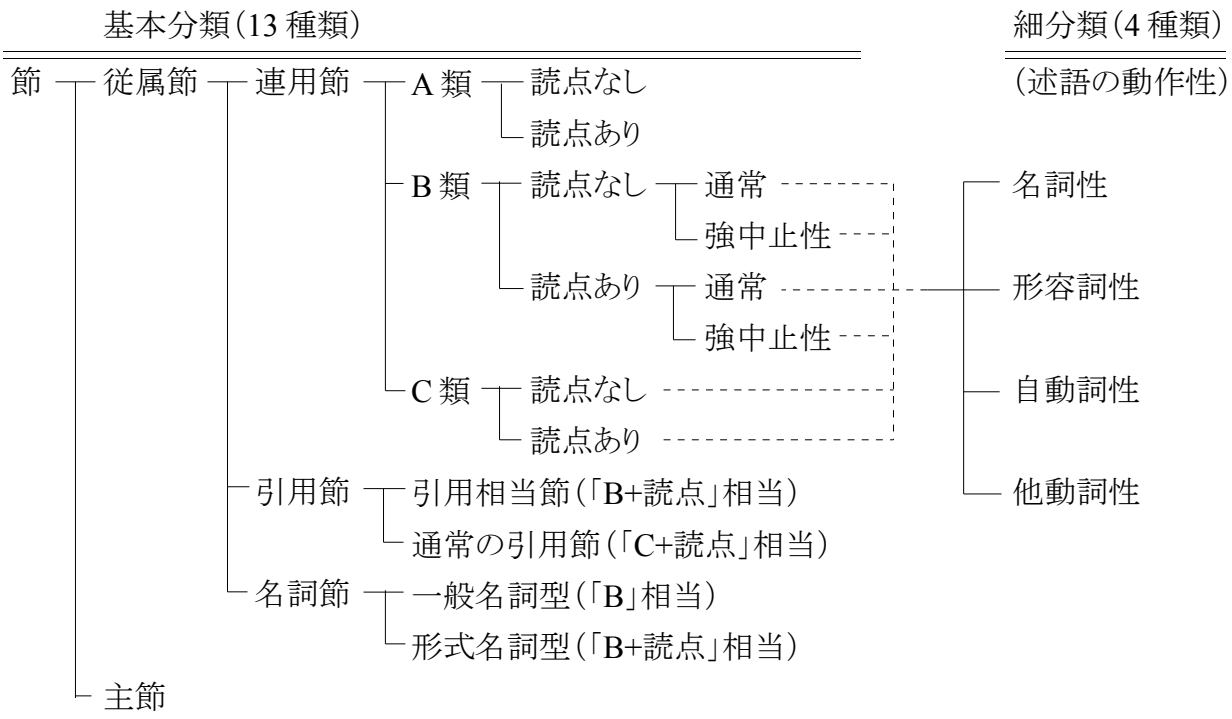


図 **2.2** 係り受け関係に着目した日本語述節の分類

まず，連用節は，A，B，C 分類と読点の有無により，6 種類に分類し，その中の使用頻度の高い B 類関連は，さらに，中止性の強さで 2 種類に分類した。引用節，連体節は，従属節の強さに応じて，連用節のいずれかの分類に畳み込むように分類した。係り受け関係を決定するうえで，以上の基本分類では足りないと見られる B 類，C 類では，細分類として，述語の動作性の強さに応じた 4 種類の分類を加えた。

2.4.2 係り受け判定規則

係り受け関係を判定する規則は，次に示す基本規則と派生規則から構成される。

［係り受け基本規則］

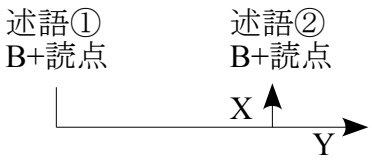
従属節 6 分類の優先順位（包含関係）に基づく規則で，次の基準で係り受け関係を絞り込む。

① 優先度の低いものは，高いものに係る。

② 優先度の高いものは，低いものに係らない。

［係り受け派生規則］

従属節 6 分類で同一の優先度の関係にある節間の係り受け関係を判定する規則で，図 **2.3** に示すように，係る条件（X），係らない条件（Y）またはその双方が記述される。述語①，②が X，Y の条件に適合しないときは，この規則は適用されない。すなわち，絞り込まれない。最終的に絞り込めない曖昧さが残ったときは,複数の候補を生成する。



X =

場合	述語①	述語②
場合 1	連用述語一般	中止型
場合 2	自動詞型	他動詞型

Y =

場合	述語①	述語②
場合 1	中止型	連用述語一般
場合 2	他動詞型	自動詞型

図 **2.3** 係り受け派生規則の例

付図に係り受け解析の例を示す。この例では，主節のほかに，4 つの従属節があるが，それらの係り受け関係は，本章の方法によって一意に決定される。

2.5 従来方式との精度比較

2.5.1 評価対象試験文

日経産業新聞からランダムに選んだ新聞記事 300 件のリード（要約または第 1 段落，計 972 文）に含まれる述語を対象に，本章の方法の効果を評価する。対象とする文における述語数の分布を表 **2.2** に示す。この表より，972 文の標本に含まれる述語数は合計 2,327 件である。係り先に曖昧さの生じる述語は，述語数が 3 以上の文の場合で，文中の後ろ側の 2 つの述語を除く述語に限られる。このことより，表 **2.2** から係り先の曖昧な述語数を求めると，全体で 661 述語である。

表 **2.2**　日経産業新聞記事における述語数の分布

述部の数	1	2	3	4	5	6	7	8	9	合計
文の数	278	320	204	100	40	18	8	3	1	972
文の割合	28.6%	32.9%	21.0%	10.3%	4.1%	1.9%	0.8%	0.3%	0.1%	100%
平均文字数	30.0	43.8	53.2	61.5	68.0	83.6	86.8	83.3	75.0	45.9

以下では，上記の新聞記事文を対象とした係り受け解析において，係り先の曖昧な述語の係り先を無作為に決定する方式，直近の述語にかける方式，従来の ALT-J/E の方式，本章の方式の解析精度を比較する。

2.5.2 無作為選択方式と直近係り先方式の場合

(1) 完全な無作為方式の場合

上記の係り先の曖昧な 661 件の述語の係り先を，無作為に決定したとしたときの正解率を考える。例えば，述語数が 5 の文の場合，無作為方式では，図 **2.4** のように，平均的に見て，係り先の曖昧な述語 3 件中，1/4+1/3+1/2＝1.083 件は，係り先が正しく決定できると考えられる。同様にして，述語数 3 以上の文の場合について計算すると，661 件中，269.8 件（40.8%）の述語は正しく決まる。逆に，それ以外の 391 件は，係り受け関係の決定に失敗することになる。以上から，無作為方式の場合，2,327 述語中，係り受けに失敗する述語の割合は，16.8%である。

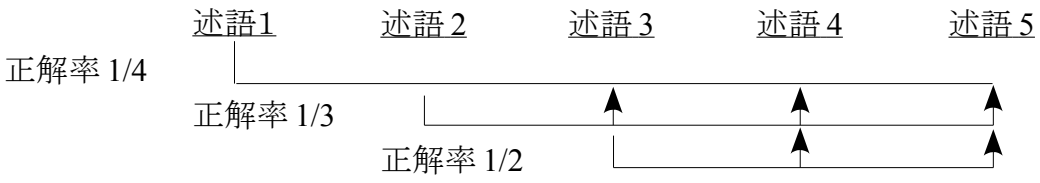


図 **2.4**　述語ごとに見た係り先の数と正解率

同様に，無作為方式で係り先を決めた場合について，文単位の述語係り受け正解率を求める。述語数 k の文の数を $n(k)$ とする。述語数 k の文の場合，述語間の係り受けの組合せ（係り受け候補）の数 L は，$L = (k-2)!$ である。すべての候補の正解率が等しいとすると，無作為に係り先を決めたときの文単位の係り受け誤り率は，$(1 - 1/L)$ である[4]。この方法で，972 文の述語間の係り受けの誤り文数を求めると，236 文となる。

(2) 係り受け交差を考慮した無作為方式の場合

特殊な場合を除いて，係り受け関係は交差しないことが知られているので，係り受け交差を禁止するという条件下で，係り先を無作為に選択した場合について，係り先選択に失敗する述語の数と係り先の誤った述語を含む文の数を求めると，それぞれ，222 述語，151 文となる。

(3) 直近係り方式の場合

経験的に，係り受け解析では，係り先が 2 つ以上あって曖昧なときは，直近に現れる述語候補を係り先とするのが良いと言われている。そこで，上記(2)の方法にさらに，この考えを加えた方法で係り先を決めたときの係り受け決定結果を正解と比較する。前と同様にして，972 文を対象に評価した結果によれば，述語単位の係り受け失敗数と文単位の係り受け失敗数は，それぞれ，189 述語，185 文であった。

(4) 読点を考慮した直近係り方式の場合

読点を持つ述語の場合は，直近の述語よりもむしろ遠くの述語に係る場合が多いことが知られている。そこで，ここでは，(3)で，読点を持つ述語は，むしろ，直近の述語に係らないとした場合について評価する。前と同様にして，972 文を対象に評価した結果によれぱ，述語単位の係り受け失敗数と文単位の係り受け失敗数は，それぞれ，112 述語，88 文であった。

2.5.3 係り受け解析精度の比較

述語間の係り受け精度を，上記の 4 方式と従来の ALT-J/E の係り受け解析方式および本章の方式の 6 者で比較した結果[5]を表 **2.3** に示す。これより，従来方式と本章の方式の関係では，次のことが分かる。

① 従来方式に比べ，本章の方式では，係り先の曖昧な述部の数が 15.3%から 2.3%に減少し，その結果，文単位に見て述語間係り受けが一意に決定できる文の割合は，73.2%から 94.4%に向上した。

4 ここでは，簡単のため，係り受け交差を認め，すべての係り受け候補が等しい正解率を持つとしている。係り受け交差を認めない方法で，1 文内の述部ごとに係り先正解率を積算する方法では，誤った係り受けの文が，本文の方法より若干（数文）減少する。

5 従来の ALT-J/E 方式の評価は，稼働中の係り受け解析プログラムを使用して行ったのに対して，本章の方式の評価は，アルゴリズムの机上トレースによって実施した。

② 述語係り先の候補として得られた第 1 位の述語が係り先として正しくない割合は，従来方式では，3.9%であったが，新方式では，0.7%に減少した。その結果，文の単位で見れば，第 1 位の文解釈候補の正解率は，91.8%から 98.4%に向上した。

また，従来言われてきたヒューリスティックスに関しても，

③ 無作為方式と直近係り先方式では，直近係り先方式の方が良いと言われているが，必ずしもそうとは言えない（表 **2.3** の第 2，第 3 の方式参照）。文頭に，独立した前提文のような従属節を持つことの多い新聞記事リード文のような場合は，直近係り方式はむしろ精度が悪い。

④ 係り受け解析精度から見れば，直近かどうかよりも，読点の扱いの方が大切と考えられる。

等が，観察される。

表 **2.3**　述語間係り受けの精度比較（標本：日経産業新聞）

集計の種別 / 係り受け決定方式		述語単位の集計（述語数合計：2,327 述部）		文単位の集計（候補文数合計：972 文）	
		曖昧さあり	誤り（注 1）	曖昧さあり	誤り（注 1）
1	無作為選択方式 1[*1]	661 件 28.4%	391 件 16.8%	372 文 38.4%	236 文 24.3%
2	無作為選択方式 2[*2]	356 件 15.3%	222 件 9.5%	260 文 26.8%	151 文 15.6%
3	直近係り選択方式 1[*3]		189 件 8.1%		185 文 18.8%
4	直近係り選択方式 2[*4]		112 件 4.8%		88 文 9.1%
5	従来の係り受け方式[*5]		92 件 3.9%		80 文 8.2%
6	本章の方式[*6]	54 件 2.3%	16 件 0.7%	54 文 5.6%	16 文 1.6%

*1　述語種別を一切考慮しない無作為方式。
*2　連用，連体，引用の述語種別を区別して，その性質を考慮した場合。また，係り受け交差を排除。その後，複数係り先候補は無作為に選択。
*3　*2 で，複数係り先候補は直近を選択。
*4　*3 で，さらに読点を考慮した場合。
*5　従来の日英機械翻訳システム ALT-J/E の方式。
*6　本章の方式。ただし，曖昧さの残った係り受けでは，直近係り選択方式 2 を採用した場合。
（注 1）文単位に出力された正解候補のうち，第 1 位の候補を対象に集計した値。

2.6 結言

長文解析の精度を低下させる大きな要因であった述語間の係り受け関係の曖昧さを解決するため，日本語の意味的な階層的表現構造に着目した，従属節間の係り受け解析方式を提案した。

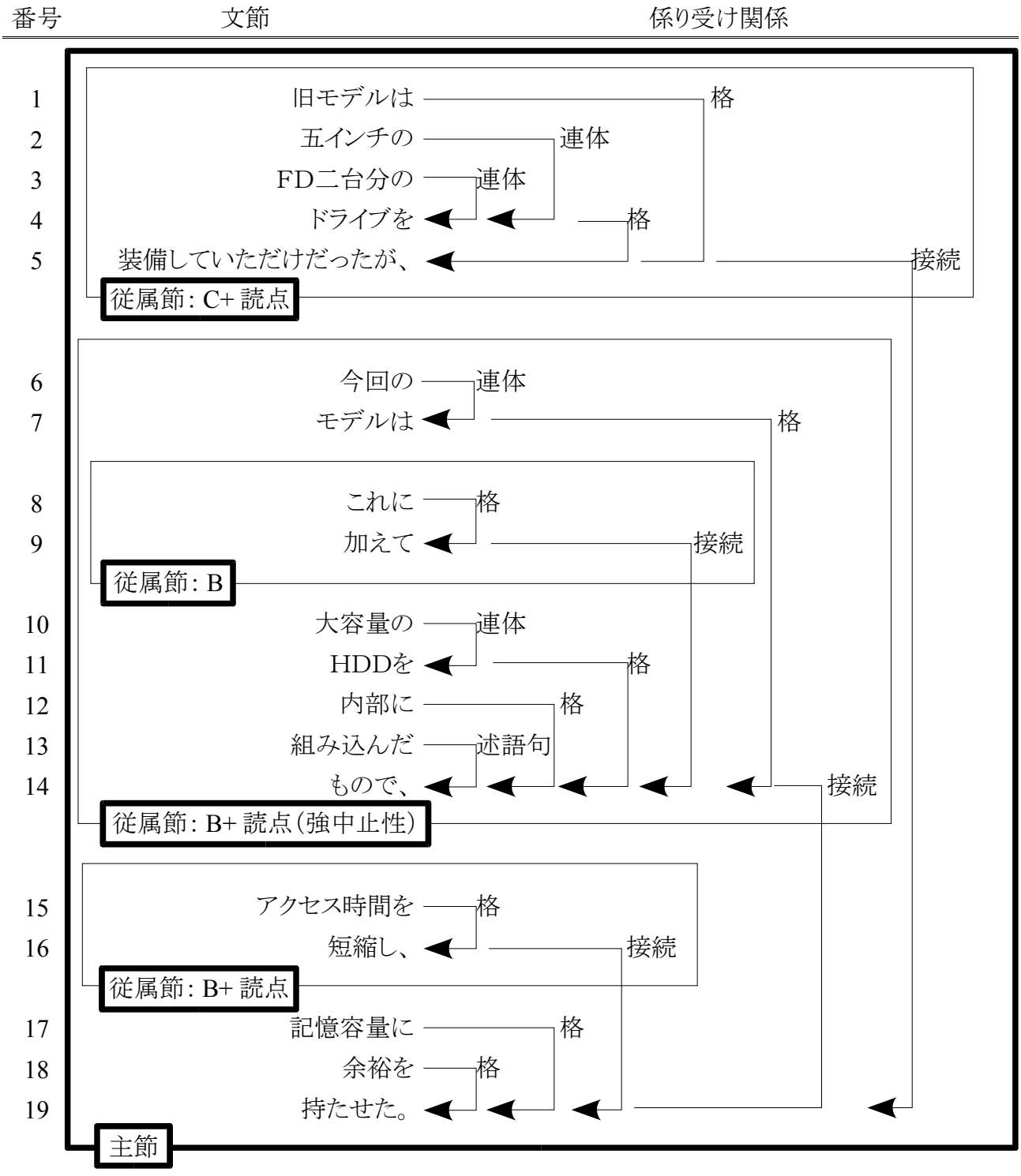
具体的には，日本語表出過程に着目した南の3段階の階層的な従属節分類を見直すとともに，さらに，その意味と形式に着目して基本分類13種，細分類4種に詳細化し，それらの係り述節，受け述節としての関係を分類整理することにより，述語間の係り受け関係を決定する方法を提案した。また，新聞記事972文（述語数合計2,327件，そのうち係り受け曖昧述語661件）を対象に，係り先の曖昧な述語の係り先を無作為に決定する方法，直近の述語にかける方式などと，従来の方法，本章の方法の解析精度を比較評価した。

その結果によれば，従来の方法では，係り先の曖昧な述語が356件残ったのに対して，本章の方法では，54件に減少することが分かった。文単位に見れば，述語間の関係が一意に決定できる文の割合は，73.2%から94.4%に向上した。係り受け関係が一意に決定できない述語に対しては，複数の係り先候補が生成され，それを組み合わせて文単位の解析候補が出力されるが，そのとき生成された文単位の第1候補の正解率は，91.8%から98.4%に向上した。

並列構造解析については，黒橋ら[17,18]によりすでに解決の見込みであることを考えあわせると，本方式によって，係り受け解析の2大問題（並列構造の解析，述語間の関係解析）がともに解決される見込みとなった[6]。

6 京都大学で公開されているKNPパーザには本章で提案した方法の一部が実装されている[19]。また，京都大学ではKNPパーザによる解析結果を人手修正したタグつきコーパスが公開されており[20]，最近，このコーパスを使用した統計的構文解析処理の研究が盛んに行なわれている。例えば，工藤らは段階的なチャンキングを用いた係り受け解析モデルに対して依存関係の有無をサポートベクタマシンを用いた機械学習より判定する方法を提案し，約2万文の学習データを用いることにより90.5%の解析精度を達成している[21]。さらに解析精度を向上させるには，解析誤りとコーパスを比較検討し，コーパスを均質化していくことが必要であると考えられる。

［入力文］
旧モデルは五インチのFD二台分のドライブを装備していただけだったが、今回のモデルはこれに加えて大容量のHDDを内部に組み込んだもので、アクセス時間を短縮し、記憶容量に余裕を持たせた。(51 単語)



付図　述語間係り受け解析の例

第 3 章　係り受け制約を利用した日本文書き替え

3.1 緒言

従来，機械翻訳において数多くの研究開発が行なわれ，翻訳業務への適用も行なわれるようになってきた[38]。しかし，依然として訳文の品質が問題となっており，新しい理論や方式の提案が期待されている[5]。

訳文品質の向上を狙って，多くの研究[39,40]が行なわれてきたが，言語が話者の対象に対する見方，捉え方をも表現する手段であることを考えると，異なる言語族間の翻訳においては，特にこの違いを克服することが重要と考えられる[41,42]。

言語間の発想の違いに着目し，もとの意味を失わないように翻訳する方法としては，第 1 に，「原文の表現や構造を分解し過ぎないよう，目的言語内で，なるべく全体の意味に該当する表現を探して置き替える方法」，第 2 に，「システムが原文の意味を変えない範囲で翻訳しやすい表現に書き替えて翻訳する方法」が考えられる。

第 1 のアプローチの例としては，言語による話者の認識の違いに着目した多段翻訳方式[43,44]などがあげられる。この方法では話者の意志や判断を示す主体的表現と対象の姿を示す客体的表現を分離して翻訳するが，客体的表現に対して，構造の持つ意味を失わないよう，構造の抽象化のレベルを設けて，段階的に翻訳している。

また，合理主義的(theoretical)なアプローチに加えて，経験主義的(empirical)なアプローチも始められ，知識ベース型の翻訳[46,46,47]や用例翻訳[48,49]等の研究が行なわれ，その効果が期待されている。用例翻訳の方法は，原文の表現を直接的に目的言語に対応させることを狙っており，やはり第 1 のアプローチの例と考えられる。

これに対して，第 2 のアプローチとしては，従来から行なわれている人手による前編集をあげることができる。前編集作業を支援するため，翻訳しやすいように言語を制限，原文をチェックするための支援プログラムを開発する試みが行なわれている[50,51]。また，変換過程を日日変換，日英変換，英英変換で構成することにより，言語間の違いを吸収する試みも行なわれている[52,53]。

言語による発想の違いは，機械翻訳しにくいところに端的に現れていると考えられるから，従来，人手による前編集の対象となっているような表現を自動的な翻訳の対象とすることができれば，訳文品質は向上すると期待できる。しかし，前編集の自動化は，無視できない副作用を生じるため，実現困難であった[7]。副作用の原因は，いわゆる同形式異内容の現象のためで，字面上は同じ表現であっても，書き替えてよい場合と書き替えてはいけない場合，または，意味によって書き替え方の異なる場合があり，自動的にその区別をすることが困難であったためである。例えば，「私は電車に乗って学校へ行く。」を「私は電車で学校へ行く。」と書き替えれば“I go to school by train.”

7 たとえば，ALT-J/E[43]では，当初，使用頻度の高い言い回しの表現を解析辞書に登録し，日本語解析のはじめの段階から使用していたが，副作用が品質向上の大きな妨げとなることが分かり，解析辞書からはすべて削除した。

という簡潔な英訳文が得られる。しかし，単純に「に乗って」を「で」に書き替えるなら，「半数は電車に乗って半数は歩いていく。」なども書き替えられてしまい，却って訳文品質の低下を招く。

そこで，本章では，書き替えの必要な現象の性質に着目し，(1)単語の詳細な文法的，意味的属性を使用して書き替え規則の適用条件を記述すること，(2)原文の解析が進行し，書き替え規則の適用条件の判定に必要な情報が得られた時点で書き替えを実行すること，の2点によって，副作用の無視できる自動書き替えが実現できることを示す。

すなわち，本章では，第1のアプローチの立場から提案されている「多段翻訳方式」のうえに，第2のアプローチの立場から，従来前編集の対象となっているような機械翻訳困難な表現や構文を自動的に書き替える方法を追加した「原文自動書き替え型翻訳方式」を提案する。具体的には，日英機械翻訳において原文自動書き替えの対象となる表現や構文の種類と性質を調べ，全体を原言語内で書き替える項目と，原言語内書き替えが困難であるため目的言語の表現に部分的に書き替えるべき項目に分け，書き替え方式と書き替え規則形式を提案する。

この方法は，訳文品質の向上を狙ったものであるが，併せて，「書き替え対象となる表現に対して，既存の翻訳機能がそのまま利用できるため，新たな翻訳アルゴリズムを作成しなくても良いこと」，「一定の表現構造を固定的に捉えることにより，構文意味解析の曖昧性が減少するため，処理速度が向上すること」などの効果[54]も期待できる[8]。そこで，新聞記事翻訳への適用実験結果に基づき，これらの効果も示す。

8 翻訳技術の発展により翻訳能力が向上すれば，翻訳困難な表現の書き替えは次第に不要となると予想されるが，構文解釈の曖昧さ減少の効果をも考えると，原文書き替えを機械翻訳システムの基本機能の1つと見なすこともできる。

3.2 書き替えの対象

3.2.1 自動書き替えの対象範囲

原文自動書き替えの対象となる表現は，以下の条件を満たす表現と考えられる。

条件 1： そのままでは適切な翻訳ができない。
条件 2： 意味を変えないような書き替え方法がある。
条件 3： その書き替えを行なえば，翻訳可能となる。
条件 4： 既存の翻訳機能に対して，悪い副作用を生じない。

これらのうち，条件 1～3 は，人手による前編集の場合と同様であるが，条件 4 は異なる[9]。

(1) 機械翻訳不能の原因

まず，第 1 の条件について考える。実際の文書で適切な機械翻訳ができない表現を分類すると，おおよそ以下のとおりとなる。

(i) 原文が間違っている。
　① 原言語の表現の約束を守っていない。（誤字，脱字，構文誤り等）
　② 表現または内容が曖昧。（解析不能）
　③ 内容が間違っている。
(ii) 既存の翻訳技術で翻訳できる範囲であるが，使用しているシステムでは能力が足りない。
　① システム（辞書，規則）のバグ。
　② 該当する表現を翻訳する機能（アルゴリズム）がインプリメントされていない。
(iii) 高度な意訳等が必要で現状では翻訳困難である。
　① 原言語の表現に直接対応する目的言語の表現がないため，話者の意図を判断して，言い直さなければならないもの。
　② 慣習の違いなどにより，訳す必要のないもの。

これらのうち，(i)は③を除き，文章校正の対象範囲であり，従来から多くの研究が行なわれている[10]。日英機械翻訳で問題となるのは，(ii)と(iii)である。

(2) 意味を変えない書き替え

次に，第 2 の条件について考える。人手による前編集の場合は，原言語内に意味を変えない別の表現が存在しなければ，書き替えはできない。これに対して，翻訳システム内部で書き替える場合は，原言語内に別の表現が無くても，目的言語に適切な表現があれば，それを直接指示す

9 人手による前編集では，着目した文の着目した表現ごとに書き替えるか否かが判断できるから，副作用のある部分での書き替えは仰止できる。これに対して，自動書き替えでは，書き替え規則に当てはまる表現すべてが書き替えの対象となるから，条件 4 は，自動書き替えの重要な条件となる。

10 例えば日本語の場合は，日本文校正支援システム REVISE[55]等が実用化されている。機械翻訳を実行する前にこれらを使用すれば，形態素レベルの誤りは，ほぼ検出訂正できる。

ることで救済することができる。

そこで，原文自動書き替えの対象を以下の 2 とおりに分類する。

(A) 着目する表現に対して，当システムで翻訳可能な別の原言語表現のある場合。
（原言語内書き替え）

(B) 別の原言語表現はないが，部分的に対応する目的言語表現のある場合。
（疑似的原言語への書き替え）

このうち(A)は，原言語内での書き替えであるため，書き替え後の文は，原言語の文としても意味の分かる文となる[11]が，(B)の書き替えは，目的言語固有の表現に対応づける書き替えであり，書き替えた後の文は，必ずしも原言語の文として意味が通じる必要はない。

(3) 書き替え後の翻訳の可否

第 3 の条件であるが，書き替えた後，翻訳可能となるか否かは，人手による前編集の場合と同様であり，実験的に確認する。

(4) 副作用のない書き替え

人手による原文書き替えでは，書き替えられる文は特定されており，他の文への副作用はない。これに対して，自動書き替えの場合は，登録した書き替え規則は該当する表現のすべてに適用されるため，書き替えてはならないものまで書き替えてしまう可能性がある。特に，原文の段階での書き替えでは，書き替え対象は字面表記で指定されることになるため，字面が一致した表現はすべて書き替えられてしまう。

これらの問題を解決するには，書き替え規則は，その適用条件を精密に記述すること，また，書き替え規則は，それぞれの規則の適用条件が判定できる情報が得られた段階で適用することが必要である。

前者の問題は，ALT-J/E の単語意味属性を使用することによって解決できると期待される[12]。また，後者の問題を解決する方法としては，構文解析の候補が出そろった時点で，書き替え規則を適用することとする。

3.2.2 書き替え対象表現の分類

既存のシステムで翻訳に失敗した表現が自動書き替えの対象候補となる。原文書き替えの規則を収集するには，翻訳に失敗した表現に対する解析結果を追跡して失敗する表現のパターンを抽出し，それに対応する翻訳可能な表現パターンを実験的に求めればよい[13]。

11 人手による前編集と同様，翻訳システムに合わせた書き替えであり，必ずしも原言語の表現として適切になるとは保証されない。

12 名詞の意味属性体系（3,000 種）を使用した 日本語用言結合価規則（約 1.3 万規則）の記述実験では，用言の場合，訳し分けの規則は十分排他的に記述できることが 判明している[47]。

13 翻訳できない表現は通常，容易に発見できるのに対して，その表現が翻訳できるような機能を新たに開発し，既存の機能と整合させるのは通常，簡単でない場合が多い。それに対して，この方法は，既存の機能に対する副作用の心配が少ない点で，改良が容易と言える。

ここでは，機能試験文（3,700 文）と新聞記事文（500 文）の翻訳実験の過程で得られた経験に基づき，書き替えによって効果の期待できる表現の種類と書き替えの方法について考察する[14]。

3.2.2.1 日本語内書き替え

原言語内書き替えの対象となる項目を以下の 3 種に分類する。ただし，日本語内書き替えが可能であっても，書き替えた後の表現が意味的に曖昧になるものは，直接英語を意識した疑似的日本語に書き替えることとし，この分類から外して次節に加えた。

(1) 縮約展開型の書き替え

動詞を共有する複数の文では，前方の動詞が省略される場合が多い。例えば，1)では，動詞の「担当する」だけでなく，助詞の「を」まで省略されているため，「米国」と「副社長」が並列に見え，助詞「は」の認定に支障が生じる。このような場合，格要素の対応関係を見て，省略された述語を補完すれば，意味解析は容易になる。

1) 社長は米国，副社長は欧州を担当する。
1') 社長は米国を担当し，副社長は欧州を担当する。

また，動詞が並列に並べられると，活用語尾か省略され，見かけ上，名詞解釈される現象が発生する。例えば，2)では，動詞「追加する」の語尾「する」か省かれているため，名詞の「追加」と解釈され，文全体の意味解析に失敗する。このような失敗を防ぐため，活用語尾を補い2')のように原文を書き替える。

2) システムが追加および削除するデータ～
2') システムが追加しそして削除するデータ～

(2) 冗長除去型の書き替え

もって回った言い方など，翻訳する必要のない表現を削除する。例えば，仕様書などでは，3)のような表現が用いられる場合が多いが，「ものである」の表現は翻訳を困難にするだけでなく，英語としてほとんど意味をなさないから削除する。

3) 既存機能を拡張することによって，システム全体の能力を高めるものである。
3') 既存機能を拡張することによって，システム全体の能力を高める。

例の 4)も同様である。通常，接続助詞「ば」は，条件接続の意味のほか，この例のように名詞の列挙を表す場合がある。条件接続か名詞列挙かを区別するには，周辺の構造と意味を広く見る必要があるため，条件接続の意味に解釈しているシステムが多いと思われる。そのような場合，例えば 4)の文も内容的に同等の表現 4')に書き替えれば問題は解決する。

14 したがって，本章で取り上げる表現は，既存のシステム（ALT-J/E）の翻訳能力を超える表現であり，システムによっては，書き替えの不要な表現も含まれると予想される。

4) 男もいれば，女もいる。

4') 男も女もいる。

(3) 構文組み替え型の書き替え

日本語の構文に直接対応する英語構文がない場合，英語に対応するよう，日本文全体の構造を書き替えてしまうもので，原文からは想像のつかないような英文を生成することができる。文脈処理[54]でも省略された主語や目的語が補完できないような場合，または，補完できたとしても適切な英文にならないような場合などに適用される。

例えば，5)の文は，「合わせる」，「生産する」の主語，目的語の双方がないため，そのままでは翻訳できない。文脈から主語，目的語を補完して訳す方法もあるが，冗長な訳文になってしまう嫌いがある。そこで，原文中のキーワード的な言葉を英語構文に対応するように組み直して，5')の形に書き替える。

5) 二機種合わせて月五百台生産する。

5') 二機種の合計月産は五百台だ。

3.2.2.2 疑似的日本語表現への書き替え

疑似的原言語への書き替え対象となる項目を以下の 3 種に分類する。

(1) 独立句的表現の書き替え

日本語の動詞性の副詞句には，英語側では単純な前置詞句に訳せるにもかかわらず，直訳すれば動詞句になり，訳文の品質が低下するものが多い。6)では「乗る」は動詞であるが，「に乗って」は手段を表す“by”に対応する意味であるので，疑似的な日本語“ニノッテ”を設け，それに書き替える。手段を表す“by”に相当する日本語として，助詞「で」があるが，「で」は多数の解析困難な多義を発生させるため使用を避け，疑似的日本語への書き替えとする。

6) 私は電車に乗って学校へ行く。

6') 私は電車“ニノッテ”学校へ行く。

なお，7)の場合も，「に乗って」が使用されているが，この場合は本動詞であるため，書き替えの対象にならない。そのため，「半数は電車に乗る。」，「残りは歩いていく。」と別々に解釈される。「乗って」と「歩いて」を共に手段として解釈させるには，すでに 3.2.2.1 節の(1)で示した縮約展開型の書き替えを適用した後，本項の書き替えを適用すればよい。

7) 半数は電車に乗って残りは歩いて行く。

(2) 様相・時制表現の書き替え

様相や時制は通常，助詞，自動詞の組み合わせ（主体的表現）によって表現されるが，名詞，動詞等によって客体化された表現で表される場合がある。例えば，8)では名詞述語「予定だ」によって「計画の意志」が示されている。また，10)は名詞述語「ところだ」によって，完了直後の状態を表している。このような表現は，8'), 10')のように客体的表現から分離し，疑似的に主体的な表現として処理するよう書き替える。

8) 山谷電気は東京に本社を移す予定だ。

8') 山谷電気は東京に本社を移す（+plan to 変形）。

9) これは私が出した予定だ。

10) バスは出発したところだ。

10') バスは出発する（+完結直後状態）。

11) 古戦場は武士が戦ったところだ。

なお，同じ名詞述語でも，9)と11)は，共に全体が「A is B」の英語構文に対応する表現であるため，書き替えの対象とならない。この区別は以下のようにして行なうことができる。すなわち，8)～11)の文はいずれも，「A は B だ」の日本語構文であるが，名詞 A と B の意味的関係を見ると，8)と10)は，A と B（「山谷電気」と「予定」および「バス」と「ところ」）が意味的につながらないが，9)と11)は，A と B（「これ」と「予定」および「古戦場」と「ところ」）の意味が意味属性体系上，上下関係にあるため，書き替えは適用せず，「A is B」の構文に訳せば良いことが分かる[15]。

(3) 接続表現の書き替え

文間の接続を表す語の中には，英語にした場合あまり意味を持たず，かえって意味不明となるような表現がある。例えば 12)では，「のに続き」は行為の順序を示すだけであるので，内部表現上，接続属性として「順序接続」を付加し，原文から削除する。

12) 高機能を追加するのに続き，改良型を導入する。

12') 高機能を追加する（順序接続），改良型を導入する。

15「は格」の名詞が，場所の意味属性を持つ場合で，「ところだ」が完了の意味となる場合の例として，「新空港は開港したところだ」等の文もある。このような場合，さらに，「は格」と動詞の意味的関係をもう一段詳しく解析する必要があるが，ここでは，今後の課題とする。

3.3 自動書き替え方式

(1) 書き替え規則の形式

書き替え規則[16]の形式を表 **3.1** に示す。本規則では，書き替えの予期しない副作用を排除するため，書き替え対象となる表現は，原文中の単語の品詞，意味属性，字面のほか文字間の係り受け関係をも使って記述される。例えば，表 **3.1** の規則を構文木で示すと図 **3.1** の上段のとおりとなる。書き替え側では，「乗り物（意味属性指定）」が「乗る（字面指定）」に対して格関係にあること，「乗る」が「行く（字面指定）」に対して接続関係にあることが条件であるが，同時に，「行く」に対しては，任意の数の要素との係り受け関係を持ってもよいが，「乗る」に対しては，「乗り物」以外の係り受けを持ってはならないことが示されている。これによって，「～に乗って～行く」の表現でも，図 **3.1** の下の例に示すように，書き替えてよい場合と書き替えてはならない場合が識別される。

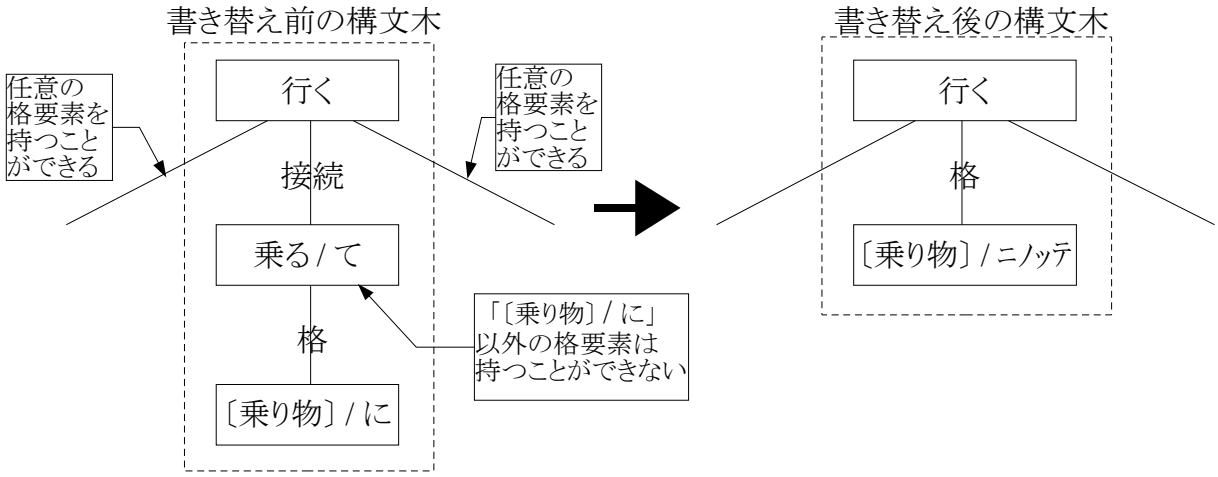
表 **3.1**　日本語書き替えルールの構成

キー単語	構文木内の位置	書き替えの対象表現			書き替え後の表現		
		構成	受け	係り	構成	受け	係り
乗る	B1	［乗り物］ +に（助詞）	任意	B2（格関係）	［乗り物］ +“ニノッテ”（助詞相当語）	＊	B3（格関係）
	B2	乗る（音便） +て（助詞）	B1	B3（接続関係）	＜削除＞		
	B3	行く［+＊］	B2	＜任意＞	＜変更無し＞	B1	＜変更無し＞

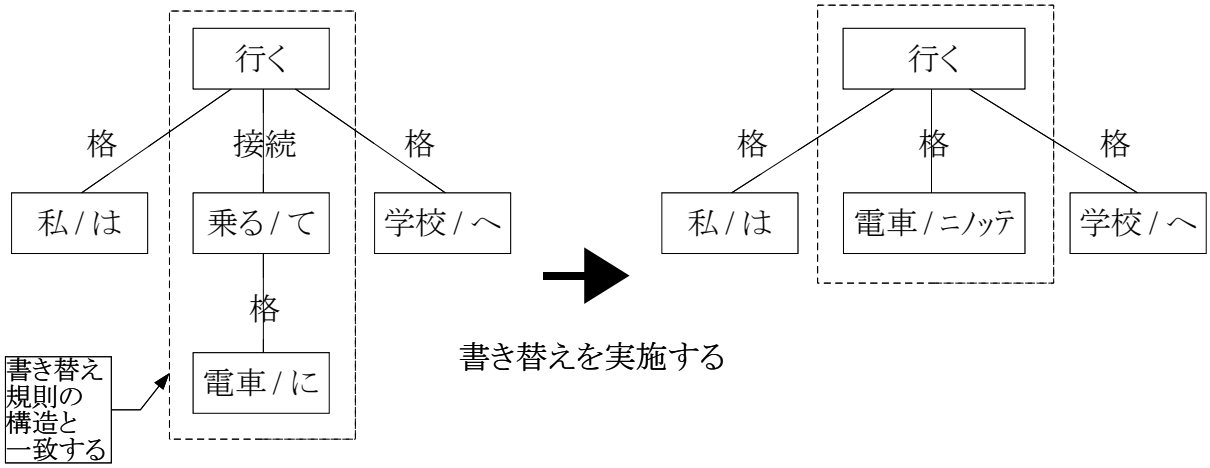
［凡例］B*n*: 部分表現（文節）の対応関係を示す。

16 生成文法で使用される「書き換え規則」と区別するため，本章では「換え(exchange)」の代わりに「替え(replace)」を使用し，「書き替え規則」とする。（英訳は例えば，「新漢英字典」（研究社）を参照。）

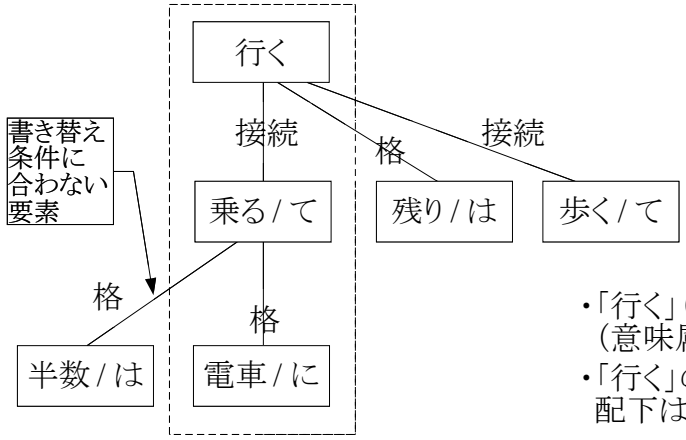
［書き替え規則］



［書き替えが適用される例］「私は電車に乗って学校へ行く。」



［書き替えが適用されない例］「半数は電車に乗って残りは歩いて行く。」



・「行く」(字面指定),「乗る」(字面指定),「電車」(意味属性指定)は書き替え条件を満たす。
・「行く」の配下のノードは任意だが,「乗る」の配下は〔乗り物〕のみでなければならない。

図 **3.1** 書き替え規則が適用される場合と適用されない場合

(2) 規則起動のフェーズ

翻訳処理は形態素解析，構文解析，意味解析などいくつかのフェーズから構成されるが，あまり早い段階での書き替えは，解析情報が不足しているため，規則の適用対象を精密に指定することが困難で，3.3.2.1(4)で述べたような悪い副作用が生じやすい。逆に，解析が深く進行した後では，後に述べるような解析多義削減効果が減少する恐れがある。

そこで，ここでは前述の規則の適用条件がチェック可能になる時点，すなわち，構文解析の直後に書き替え規則を起動することとする。図 **3.2** に，書き替え処理の位置と書き替え処理の構成を示す。

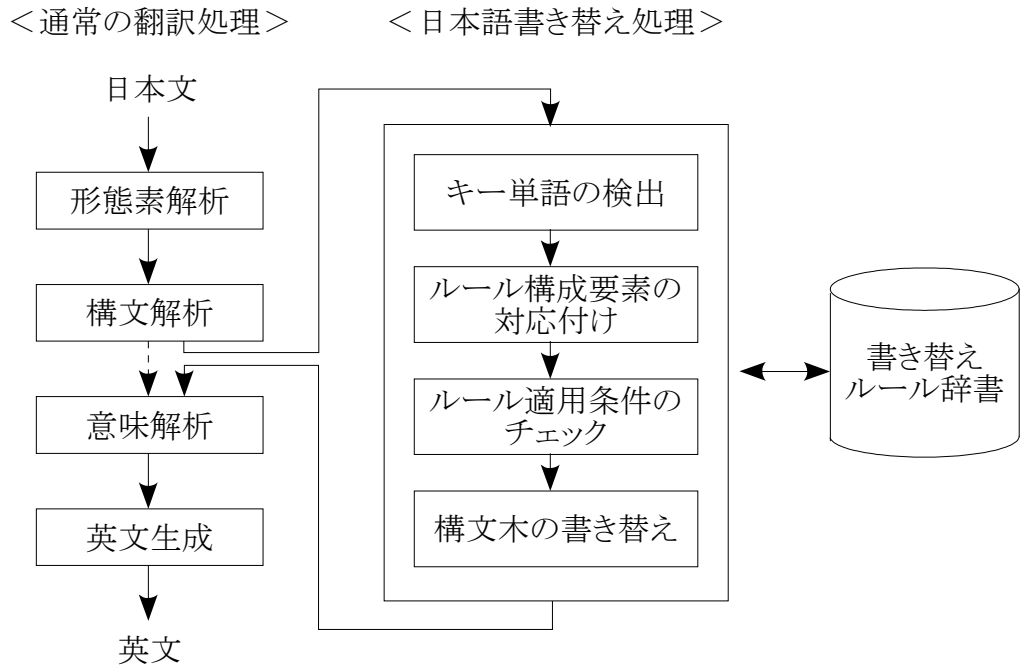


図 **3.2**　日本語書き替え型翻訳方式の構成

(3) 構文多義の扱い

構文解析では，構文上の多義は解消せず，いくつかの解析候補が残ることが多い。したがって，同一の原文に対する解析結果でも，書き替え規則が適用可能なものと適用不可能なものが生じることがある。その場合，両者を比べると，適用する知識内容の違いから，以下の理由で，書き替え規則の適用できる解釈候補の方が，相対的に正しい解釈になっていることが推定される。

① 構文解析では，単語の品詞や文節の種類などの文法的知識が使用されるのに対して，書き替え規則では(1)で述べたように，単語の意味属性等の意味的知識などが使用される[17]。

17 ここでは，むしろ，文法的知識の範囲で文構造を解析する技術を「構文解析」と呼んでおり，単語の意味属性等を扱う解析を「意味解析」と呼び，構文解析と分けて考えている。したがって，本章の書き替え方式は，従来の構文解析に，一部意味解析を追加する方法となっている。

② 構文解析では，文節間の関係が 2 項関係を基本に解析されるのに対して，書き替え規則では，多項関係で捉えられるため，表現構造の持つ意味が捉えやすい[18]。

例えば，前節 1)の例文では，構文解析の結果は，図 **3.3** に示すような 2 つの解析多義を持つが，[解釈 1]には表 **3.1** の規則が適用できるのに対して，[解釈 2]には適用できない。この場合，適用できない解釈の方を単に削除することにより，解釈は一意に定まる。

＜日本語原文＞私は電車に乗って学校へ行く。

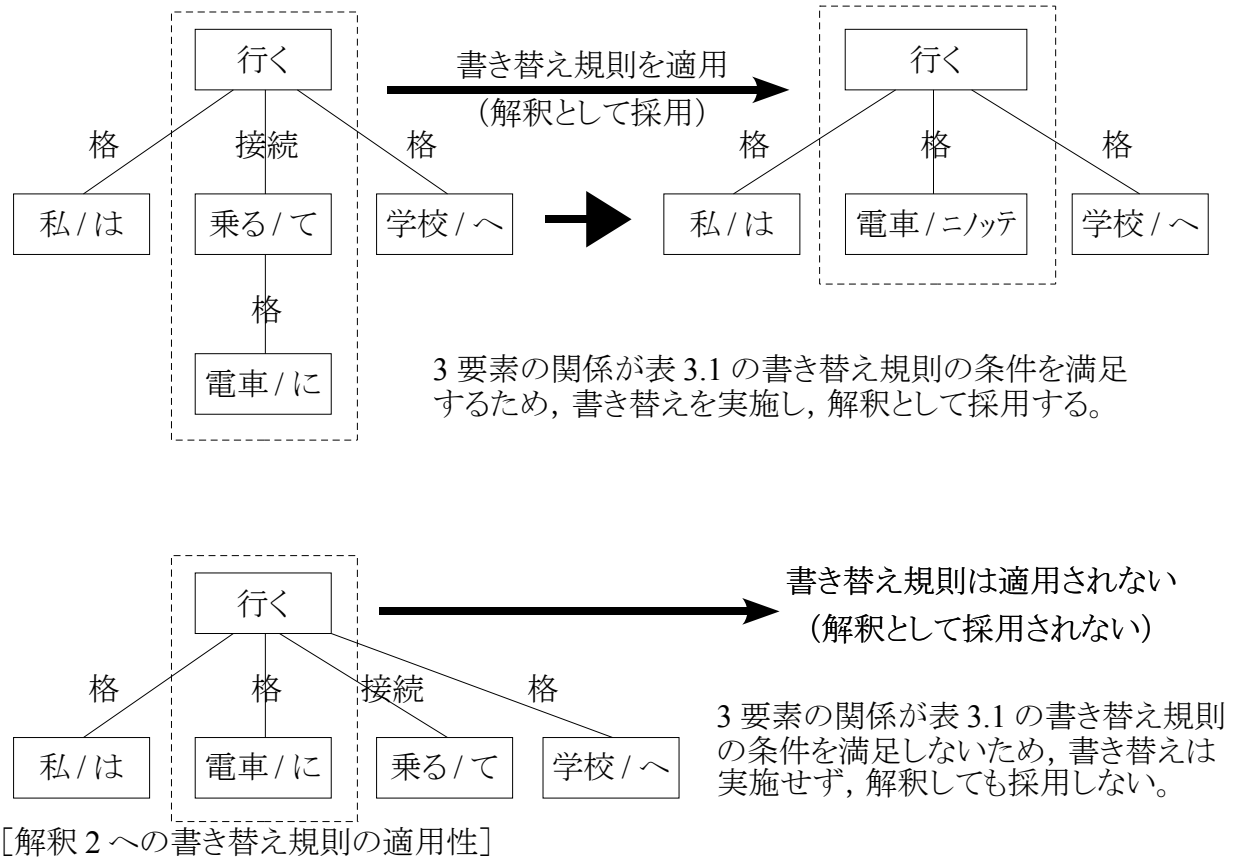


図 **3.3**　書き替えによる多義削減の例

18 ここでは，構文解析の手法として，係り受け解析を前提としている。3 項以上の閨係から句構造を決定するような構文解析の場合は，書き替えの効果は①のみとなる。なお，ALT-J/E システムでは，意味解析や日英変換の過程で，3 項以上の文要素の意味的関係から，構文絞り込みや構文変換を実施しているが，本章の方法は，独立した書き替えのプロセスを設け，書き替えの内容に応じて，可能な限り早い位置で書き替えを実行することを主眼としている。

3.4 実験と評価

3.4.1 実験と評価の条件

3.2 節, 3.3 節で述べた日本文書き替え方式を, 日英機械翻訳システム ALT-J/E のうえにインプリメントし, 日本文自動書き替えを実施した場合としない場合について, 比較評価を行なった。

(1) 対象試験文と実装した規則数

日経産業新聞の 32 記事のリード文 102 文を翻訳対象とした。原文の文平均の文字数は 40.2 文字/文, 単語数は 21.2 単語/文である。各記事のリード文は 3〜5 文から構成されており, 文脈を持っているため, 記事単位に翻訳する[19]が, 評価は文単位に行なう。

また, 書き替え規則は, 上記の試験文を含む新聞記事 500 文と日英翻訳システム用の機能試験文(第 2 版 3,700 文)[41]の翻訳実験に基づいて作成した 940 規則を実装した。

(2) 訳文品質の採点基準

訳文品質の採点基準は, ALPAC[57]の 9 段階採点基準を以下の観点で見直した 10 点満点法を使用した。

① 訳文だけで原文の意味が理解できるものを 6 点以上とし, 合格とする。
② 簡単な後修正で使える英語となる文を 8 点以上とし, 秀訳とする。

なお, 採点は, 翻訳会社の 3 名の日英翻訳家がお互いに独立に行ない, その平均点の小数点以下を四捨五入した整数値を訳文の得点とした。

3.4.2 実験結果と考察

日本文自動書き替え実験の結果を表 **3.2**, 表 **3.3** に示す。試験に使用した 102 文に対して, 書き替え規則の適用された文は, 44 文(43%)で, 延べ適用箇所は 52 箇所であった。付表に, 自動書き替えを適用しない場合と適用した場合の翻訳結果の例を示す。

以下, 規則の適用された文における訳文品質の変化と意味解析多義の変化について考察する。

19 記事内の文脈から省略された主語と目的語を自動的に 補完して翻訳する[56]。

表 **3.2**　書き替え前後の得点変化

■:品質低下領域

前＼後		書き替え後の得点												
		不合格点						合格点					平均	
	点	0	1	2	3	4	5	6	7	8	9	10	4.3 点	
書き替え前の得点	不合格点 0													35 文 80 %
	不合格点 1	■								1			1	
	不合格点 2	■	■						1	1			2	
	不合格点 3	■	■	■	4	1	2	5			1		13	
	不合格点 4	■	■	■	■	1	2	2	3		1		9	
	不合格点 5	■	■	■	■	■	1	3	4	1	1		10	
	合格点 6	■	■	■	■	■	■	3	1	2			6	9 文 20 %
	合格点 7	■	■	■	■	■	■	■	2				2	
	合格点 8	■	■	■	■	■	■	■	■			1	1	
	合格点 9	■	■	■	■	■	■	■	■	■				
	合格点 10	■	■	■	■	■	■	■	■	■	■			
	平均 6.7 点				4	2	5	13	11	5	3	1	合計 44 文	
		11 文　25%						33 文　75%						

［備考］対象試験文は新聞記事 102 文（32 記事）で，
文平均の文字数は 40.2 文字/文（21.2 単語/文）。

表 **3.3**　書き替えルールの適用箇所と訳文品質向上効果

書き替え種別	番号	書き替えのタイプ	ルール適用箇所	訳文品質向上効果	合格文数の増加	訳文コンパクト効果
日本語内書き替え	1	縮約展開	7 箇所（7 文）	1.7 点	1 → 5	+ 1.3 語
	2	冗長除去	2 箇所（2 文）	3.5 点	0 → 2	- 0.9 語
	3	構文変換	12 箇所（11 文）	1.6 点	3 → 5	- 0.1 語
疑似的日本語表現の書き替え	1	独立句的表現	21 箇所（19 文）	2.3 点	3 → 15	- 1.6 語
	2	様相時制表現	7 箇所（7 文）	2.0 点	2 → 6	- 2.3 語
	3	接続表現	3 箇所（3 文）	1.7 点	1 → 3	±0.0 語
計又は平均	-	----	52 箇所（44 文）	2.0 点	9 → 33	- 0.8 語

［備考］(1) 対象試験文は 102 文で、文平均の文字数は 40.2 文字/文（21.2 単語/文）。
(2) 複数の書き替えルールの適用された文が 10 文あるが、書き替え効果は、適用されたルールごとに調べて集計した。

(1) 訳文品質向上効果

規則の適用された 44 文のうち，33 文（全体の 32%）において訳文品質向上効果がみられた。全体の訳文合格率は 55%から 79%に向上した。本方式適用前後の得点分布を図 **3.4** に示す。

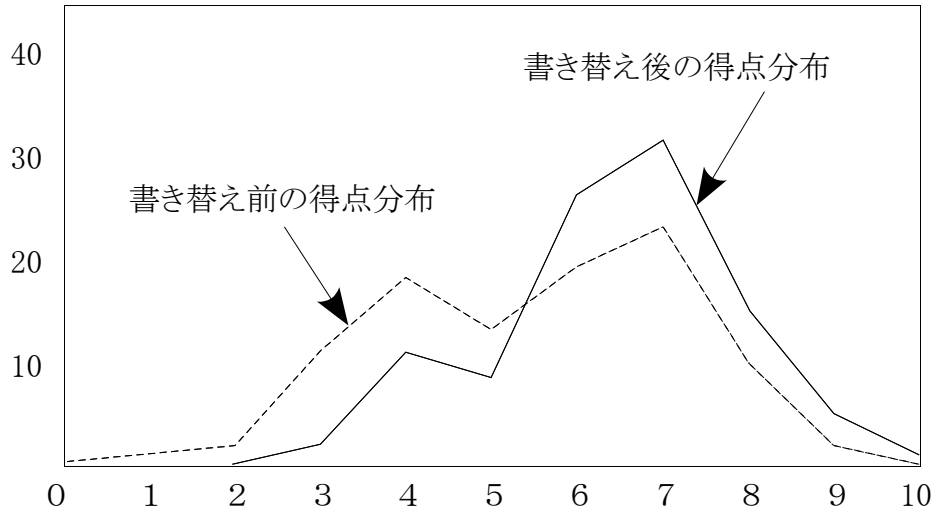


図 **3.4** 書き替え方式による訳文品質向上効果

102 文全体の平均点は適用前の 5.7 点から 6.6 点に向上したのに対して，規則の適用された 44 文の平均点は，適用前の 4.3 点から適用後は 6.7 点となり，平均 2 点以上向上した。

特に，書き替え前の翻訳結果が 4～5 点の文の場合，その多く（15/19＝79%）が，6 点以上の合格点となった。元の点が 3 点以下の文では，書き替え対象外の誤りの影響が大きいが，それでも合格点まで向上した例（9/16＝56%）はかなりあった。

規則適用によって不合格（5 点以下）から合格（6 点以上）に変化した例文は 24 文であるが，その内訳は，日本語内の書き替えによるもの（5 文），疑似的日本語への書き替えによるもの（18 文），それらの両者によるもの（1 文）であり，疑似的日本語への書き替えの方が効果が大きい。

疑似的日本語への書き替えは，後の英文生成処理への負担が減少し，書き替え後の翻訳誤りの発生を防ぎやすい等の利点もある。今後，さらに強化していきたい。

書き替え規則のタイプとその効果の関係を見ると，独立句的表現の書き替え規則の適用例が最も多く，訳文品質向上効果も大きい。

(2) 訳文コンパクト化の効果

訳文のコンパクト化の観点からみると，縮約展開型書き替えでは，必然的に訳文の単語数が増加する（平均 4.3 語増）が，その他の書き替えでは減少している（平均 1.8 語減）。全体としてみれば，訳文の単語数の減少は，文平均 0.8 語程度にとどまっており，訳文コンパクト化の効果はあまり期待できない。

(3) 解析多義削減効果

規則が適用された 44 文の意味解析の多義は，平均 5.4 から 1.3 に減少した。この現象は，上記の訳文品質向上効果を生んでいると同時に，意味解析処理の高速化にも役立っている。

3.5 結言

機械翻訳の品質を向上させるための 1 つの方法として，(1)精密な単語意味属性を使用して書き替え規則を記述すること，(2)書き替え規則適用条件の判定可能な情報が得られる構文解析結果に規則を適用すること，によって副作用の少ない原文自動書き替え型の翻訳方式を実現した。

書き替える原文対象は，①着目する表現に対して，当システムで翻訳可能な別の原言語表現のある場合（原言語内書き替え方式）と，②別の原言語表現はないが，部分的に対応する目的言語表現のある場合（疑似的原言語への書き替え方式）の 2 つに分け，合わせて 6 種類の自動書き替え項目を実現した。

新聞記事を使用した翻訳実験結果によれば，書き替え規則の適用された箇所は 102 文中，44 文，延べ 52 箇所であった。そのうち訳文品質向上効果のあった文は 33 文である。また，適用された文の構文意味解析の多義の数が平均 5.4/文から 1.3/文まで減少した。実験の結果，本方式は，翻訳品質向上，多義解消の双方において大きな効果があることが分かった。

また，本方式はインプリメントの観点からみても，「翻訳困難な表現の翻訳に，既存の翻訳機能がそのまま利用できる」点で，大きな利点があり，今後の訳文品質向上策として有望であると判断できる。

今後の課題としては，節や構文全体の書き替えへの拡張が考えられる。その際，本章で示した疑似的日本語への書き替えによって入力文を目的言語の表現に接近させれば，一種の翻訳バイパスができることになるので，ハイブリッド型の翻訳方式になる。また，書き替えのタイミングの問題では，本章は構文解析の後の書き替えを示したが，形態解析後の解析誤りを回復するための書き替えなどについても適用可能であり[58,59]，翻訳品質の向上には有効である。

付表　原文自動書き替えによる訳文変化の例（**1/2**）

種別	番号	原文	書き替え対象部分	自動書き替えを適用しないときの翻訳結果	自動書き替えを適用したときの翻訳結果
日本語内書き替え	1	二階にショールーム、三階に商談室、会議室、セミナー室を開設し、事務室は四階以上になる。	［縮約展開］ ショールームを開設し、	*C.Ito Techno-Science Corp.* will set up conference room, a meeting room and a seminar room in the second floor to a show room and the third floor and an office will reach the fourth and higher floors. ＜評価＝3 点＞	*C.Itoh Techno-Science Corp.* will set up a show room in the second floor and will set up conference room, a meeting room, and a seminar room in the third floor and an office will reach the fourth and higher floors. ＜評価＝6 点＞
	2	同社がこの二十五年間で蓄積してきた自治体向けのアプリケーションプログラムを集大成するもの。	［冗長除去型］ 集大成するもの。 → 集大成する。	Sales of local governmcnt specialized system with an office computer is the thing which the application program of the aimed at local governments that this company has stored in these 25 years is compiled. ＜評価＝2 点＞	It will compile the application program for the local government that this company has stored in these 25 years. ＜評価＝7 点＞
	3	二機種合わせて月四百台生産する。	［構文組み替え］ → 二機種の合計月産は *400* 台だ。	It produces *Midori Denki* 合わせて in 2 models in 400 units per month. ＜評価＝2 点＞	The monthly output of 2 models is 400 units. ＜評価＝8 点＞

注）翻訳結果の斜体字：文脈処理で，記事内の他の文から補完された要素を示す。

付表　原文自動書き替えによる訳文変化の例（**2/2**）

種別	番号	原文	書き替え対象部分	自動書き替えを適用しないときの翻訳結果	自動書き替えを適用したときの翻訳結果
疑似的日本語表現の書き替え	1	ソフト会社、N&Cソフトウエアはシステムハウスの ユニコムオートメーションと共同でパソコンを使った カラー印刷システムアトリエ・ビットを開発した。	［独立句書き替え］ と共同で → jointly with を使った → using	N&C Software Corp., a software company, developed Atlier Bit, the system of color printing that used a personal computer by Unicom Automation Corp. and the synergic of a system house. ＜評価＝4 点＞	N&C Software Corp., a software company, developed Atlier Bit, the color printing system using a personal computer, jointly with Unicom Automation Corp., a system house. ＜評価＝9 点＞
	2	富山センターはソフト開発要員五十人でスタート、百五十人に増やす計画。	［様相時制書き替え］ 計画。 → be planning to ～	The Toyama Center is a start in a development staff of 50 and is a plan increased in 150 person. ＜評価＝4 点＞	The Toyama Center starts in a development staff of 50 and is planning to increase the Toyama system Center to 150 people. ＜評価＝7 点＞
	3	出版取次はもともと利益率が低いことに加えて、出版物も需要が鈍化しているため苦しい経営を余儀なくされている。	［接続表現書き替え］ ことに加えて、 → not only ～ but also ～	Because it adds a puhlication agency to that a profit rate is low originally and the demand for publication is slackening, tight management is made to be unavoidable. ＜評価＝4 点＞	Because not only the profit rate of a publication agency is low originally, but also the demand for publication is slackening, tight management is made to be unavoidable. ＜評価＝7 点＞

第 4 章　日英機械翻訳に必要な結合価パターン対

4.1 緒言

機械翻訳において意味解析の重要性が指摘されている。意味解析の方法としては，単語の共起関係に着目して単語相互の意味を決定する方法が研究されているが，中でも，動詞の意味解析においては，動詞と名詞の意味的な共起関係に着目した結合価パターンを使用する方法が有効であることが知られている。この方法を実現するには，パターンの記述精度の問題とパターン対収集方法の問題がある。

パターン記述精度の問題については，日英機械翻訳の場合，格要素となる名詞の意味属性を約 2,000 種類以上の分解精度で分類すれば，慣用表現を除き，日本語の動詞を訳し分けられるようなパターン対が記述できることが知られている[47]。

これに対して，パターン対収集の問題についても，すでに，種々のヒューリスティックスや学習技術を応用した方法が提案されている。しかし，どれだけのパターン対を作成すればよいか不明であること，網羅的にパターン対を作成するのに必要な用例を実際の文書から収集するのは困難なことなど，様々な問題があり，実用できるレベルにない。例えば，黒橋らは例文とシソーラスを用いて文型を同定する方法を提案した[60]。また，アルモアリムらは自動学習の手法を用いた翻訳ルールの自動抽出方法を提案し，6 動詞に対し各 27～80 の対訳用例を用いた抽出実験に成功している[61,62]。しかし，これらの方法を使用するには，学習に必要なだけの種類と量の例文を入手できることが前提となる[20]。例えば，日英翻訳の場合，使用頻度の高い和語動詞のパターン対をほぼ網羅的に学習させるには 1,000 万ペアの日英対訳文が必要であると言われている[63]。しかも，その対訳文は動詞と名詞を組み合わせた単純な文形式で与えられなければならない。実際の文書から得られた例文は通常，複雑な構造を持つ場合が多いので，目的にあわせて単純化する作業が必要となる。このように，膨大な量の単純化された用例を実際の文書から機械的に収集することは，事実上，不可能である。

これに対して機械翻訳では，一度網羅的なパターン対が完成すれば[21]，繰り返しパターン対を作成する必要はない[22]。また，訓練されたアナリストによれば，適切な対訳用例があれば，類推能力によって，1 用例から 1 パターン作成することができると推定される。これらの点を考えれば，現状では，パターン対作成作業はむしろ人手を中心に進め，計算機はあくまで作業支援に使用す

20 自動学習の方法では，モデルを単純化せざるを得ないなどの理由で精度上も問題があり，実用に展開するのは困難とみられる。

21 種々な言語現象で見られるような使用頻度の低いパターン対でも，それを合計した出現頻度は無視できない程度となることが予想されるため，機械翻訳システムにおいては（専門分野依存は別にして）一般的なパターン対はあらかじめ 網羅的に整備する必要がある。

22 ここでは専門分野固有のパターン対を除く。専門パターン対については後で触れる。

るのが現実的と考えられる[23]。

そこで，本章では，人手によるいくつかのパターン対作成の方法について部分的な作業実験をし，その結果から，日英機械翻訳ではどれだけの数のパターン対が必要か，また，それは実際にはどのような方法によれば作成できるかを明らかにする。

具体的には，単語当たりの語義数が多いためパターン対の相互関係が問題となる和語動詞約1,000 語の中の代表的な動詞を対象に，(1)人間用の和英辞書に記載された語義に着目する方法，(2)日本語の語義に着目する方法，(3)人間の知識を内省して用例を作成し，その用例からパターン対を作成する方法の 3 種類の方法を示し，それらの方法でどれだけのパターン対が収集できるかを検討する。また，得られたパターン対の数から，和語動詞全体では最終的にどれだけの数のパターン対を作成すればよいかを推定し，その作成方法について議論する。

最後に，漢語動詞，形容詞系述語，名詞述語のほか，用言性の慣用表現を含むパターン対全体で必要なパターン対の数についても考察する。

23 計算機による支援としては，不足しているパターン対の作成支援のほか，人手で作成されたパターン対の相互無矛盾性の検証支援も重要である。自動学習技術の応用研究には，パターン対と単語意味属性体系の間の相互矛盾を論理的に検証する 仕組みの研究が期待される。

4.2 前提条件

4.2.1 パターン対記述の枠組み

機械翻訳において用言と名詞の共起関係の知識を結合価パターンにまとめるには，対象となる用言の種類，名詞の意味分類の方法等が問題となる。特に，名詞の意味分類では，翻訳する言語ペアによって必要とされる分解能に差が生じる。日英機械翻訳の場合は，日本語の用言と英語の用言の意味的な対応関係が記述できる程度の分解能を得るため，日本語の名詞の意味を2,000種程度以上に分解整理することが必要とされている[47]。本章では，この条件を満たしていると見られる日英機械翻訳システムALT-J/E[44]の枠組みを用いてパターン対の作成方法を検討する。以下では，ALT-J/Eのパターン記述の枠組みを示す。

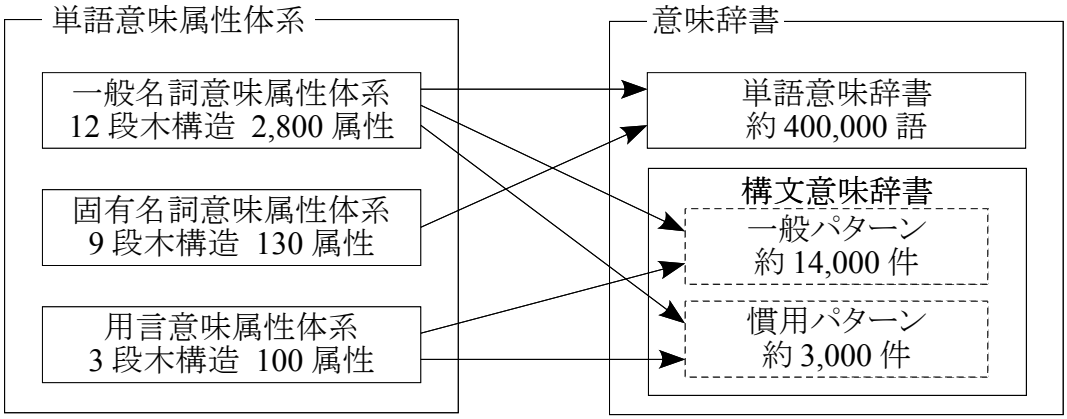


図 **4.1** **ALT-J/E** におけるパターン対記述の枠組み

ALT-J/Eの結合価パターン記述の枠組みは，図 **4.1** に示すように，日本語名詞に対する単語意味属性体系と2つの意味辞書（単語意味辞書，構文意味辞書）から構成される。単語意味属性体系は，2種類の意味属性体系から構成されるが，結合価パターンの記述には，そのうちの一般名詞意味属性体系が使用される。これは，日本語名詞の意味的な用法を約2,800種の属性名で表現し，それらの相互の意味的関係を12段の木構造に整理したものである。単語意味辞書では，単語約40万語の持つ意味（1単語1つ以上）が単語意味属性を用いて記述されている。また，構文意味辞書は日本語の結合価パターンとそれに対応する英語の構文パターンをペアとして持つ。これらの辞書は，構文解析結果の絞り込み，動詞の訳語の選択，名詞訳語の選択等の意味解析に使用される。

ALT-J/Eの結合価パターンは，用言（動詞，形容詞），格要素（主名詞＋助詞），副詞要素，様相情報から構成される。主名詞は，通常，日英の動詞が訳し分けられる最小限の深さの意味属性を用いて記述される[47]。意味属性で代表できないような名詞の場合は，名詞そのものが使用され

る。格要素の主名詞が意味属性によって指定されたパターンを一般パターン，1 つ以上の格要素の主名詞が名詞そのものによって特定されたパターンを慣用パターンと呼ぶ[24]。慣用パターンは，慣用表現や固定化した比喩的な表現に対する日英間の対応付けのために使用される。本章では，一般パターン対の収集を対象とする。

結合価パターンは，述語となる用言（動詞，形容詞）ごとに作成される。日本語では名詞が述語になる場合があり，この「名詞＋だ（です）」型の述語は一般に英語では名詞補語として訳出されるが，名詞補語には訳出できないものに対し名詞を述語とするパターンが作成される。例えば，「今日は晴れだ。→ It is fine today.」や「あなたに質問です。→ I ask you a question.」などである。また，述語が複合語の場合，例えば，「成功は努力次第だ。→ Success depends on one's efforts.」に対しても同様にパターン対が作成される。

4.2.2 パターン対作成の方法

精度の良いパターン対を効率的に作成するには，対訳用例からパターン化すべきものを発見しパターン対の作成を支援する仕組みと，作成したパターンと既存のパターンとの間の無矛盾性を検証する仕組みが大切である。ALT-J/E では，パターン対作成の過程を支援するために図 **4.2** に示すような仕組みを実現した。

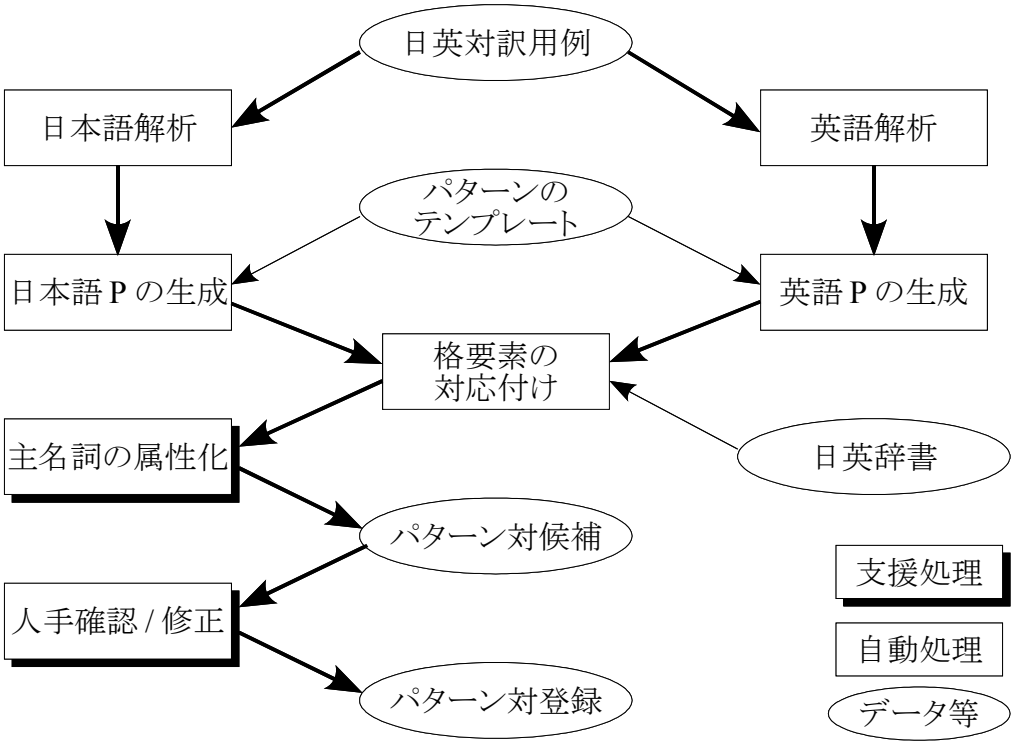


図 **4.2** **ALT-J/E** におけるパターン対作成支援の仕組み

24 一般パターン，慣用パターンのほかに，特定の専門分野を対象とする専門パターンがあるが，本章では専門パターンは扱わない。

(1) パターン対作成支援の方法

日英機械翻訳用のパターン対の構造はその大半が約 10 種類のテンプレートで記述できることが知られている[64]。したがって，これを使用して与えられた日英対訳用例の中から日本語側，英語側のパターン要素を指定すれば，最低限のパターンは容易に作成できる。しかし，質が良く汎用性の高いパターン対を作成するには，パターンの適用範囲を決定する名詞要素の記述が大きな問題となる。この作業を支援するため，ALT-J/E では以下のコンピュータ支援処理を実現した。

例えば，用例「彼は電話を引いた。→ He installed a telephone.」に対して，まず「X〔主体〕が/“電話”を/引く → X install a telephone」というパターンが作成される。支援処理は単語意味辞書を見て名詞“電話”の意味属性とその上位の意味属性を表示するので，アナリストはこれを見て“電話”の部分を汎用的な意味属性に置き換えてパターンを作成するか，そのまま辞書に登録するかする。そのまま登録した場合，その後，日本語動詞“引く”，英語動詞“install”である用例が追加されたとき，支援処理が再度，ヲ格の名詞（複数）に共通する意味属性を表示するから，それを見てアナリストはパターンを汎用化できる。用例が増加すれば意味属性候補の判断はより正確になる。

(2) パターン対相互チェック支援の方法

結合価パターンは述語を見出し語として登録されるから，見出し語が異なるパターンの間で相互に干渉することはない。したがって，パターン相互の無矛盾性をチェックするには，同一の見出し語を持つ用例を対象に翻訳実験を行えばよい。そこで ALT-J/E では，パターン相互の無矛盾性チェックを支援するため，以下の手順を半自動的に実施する仕組みを実現した。まず，前述の(1)の処理のあと，パターン作成に使用した用例とそれに対する機械翻訳の結果を保存する。再び(1)の手順で新パターンを作成したときは，新パターンを暫定的に登録した後，同一の見出し語を持つ既存の用例を対象に翻訳実験を行う。その結果を過去の翻訳結果と比較して，差分の生じた用例とその翻訳に使用されたパターン対を出力する。アナリストはそれを見て，新パターンの最終的な登録の可否を判断する。

無矛盾性チェックの結果によっては，新パターンの作成だけでなく，既存パターンの修正が必要な場合もある。パターンの修正はまた(1)に戻って実行される。

4.3 パターン対収集の方法

上記の支援システムはあくまで人手作業を支援するものであり，すべての知的判断は人手で行われる。そして，判断に使用される基本情報は日本語用言の語義もしくはその用例である。そこで，用言の語義および用例の入手方法に着目して，パターン対収集の手順を3段階に分けて考える。すなわち，(1)和英辞書の語義を参照する方法，(2)日本語の語義に基づく方法，(3)人の知識を内省する方法の3種類の方法を順に適用してパターン対を作成するものとする。

4.3.1 和英辞書の語義分類に基づく方法

(1) パターン対収集の方法

パターン対を収集する第1の方法として和英辞書の情報を参照する方法を考える。人間用の和英辞書には，日本語の用言に対して，語義とそれに対応する英語の動詞や語法，例文などが記載されている。したがって，これらの辞書に記載された語法や例文を分析し，格要素，副詞要素などの日本語側の制約条件を整理すれば，日本語動詞と英語動詞のペアに対してパターン対を作成することができる。例えば，ライトハウス和英辞典[65]には，動詞「上がる」に対して5つの語義が示され，第2の語義の例文として次の文がある。

彼の学校の成績が上がった。
His school record *has improved*.

この例文の文要素を分析し，若干の情報追加を行えば，図 **4.3** のようなパターン対が得られる。

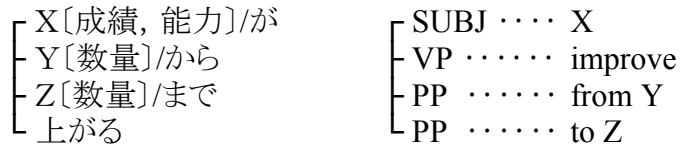


図 **4.3** 人間用の辞書を使用したパターン対作成の例

本章では，何冊かの和英辞書[25]を使用してパターン対を作成した。

(2) 収集されたパターン対の数

和英辞書に含まれる主な用言5,600語に対して上記の方法でパターン対を作成した。得られたパターン対は，当初，一般パターン10,000件，慣用パターン5,000件であった。その後の見直しにより，一般パターンの中に統合できるものが含まれていること，また，慣用パターンの中にも汎用化できるものがあることなどが分かり，辞書から収集したパターン対は一般表現10,000パターンと

25 慣用パターン作成では，一般辞書のほかに慣用表現辞書も使用した。

慣用表現 3,000 パターンとなった。

(3) 翻訳実験での充足性

上記で得られたパターン対を使用して，情報処理装置関連の仕様書（1,361 文）の翻訳実験を行った。その結果によれば，試験文中に現れた用言の種類は 142 件，翻訳に必要なパターンは 201 件であるのに対して，本節の方法であらかじめ準備できていたパターン対は 120 用言に対する 154 件であった。試験文中の 22 の用言（22 パターン）はパターン対が登録されていないこと，また，23 の用言に対しては合計 25 のパターンが不足していることが分かった。

この例から見れば，用言数で 15%（22/142），パターン数で 23%（22+25/201）が不足していることになる。中でも，パターン対が不足している用言は，単語当たりの語義の多い和語動詞が多い。

4.3.2 日本語辞書の語義分類に基づく方法

(1) パターン対用例収集の方法

前節で見たように，和語動詞は語義が多いため，通常の和英辞書の語義分類だけでは翻訳パターンを網羅的に収集することは困難である。これに対して，和語動詞については，かねてより日本の言語学者（20 名あまり）を中心にその語義と対応する用例を収集分析する研究が進められており，すでに 861 動詞に対して語義と語義ごとの用例（ただし，日本語用例のみ）が IPAL 動詞辞書[66][26]としてまとめられている。そこで，本節では，第 2 の方法として，日本語の語義をより詳細に分類する立場から，この辞書の用例を使用したパターン対の収集を考える。

具体的には，IPAL 動詞辞書の各語義に示されている用例に対して，日本語原文に忠実で，かつ，英語としても十分通用する英訳文を翻訳家に作成してもらい，その対訳データからアナリストがパターン対を作成する方法でパターン対の収集を試みる。

(2) 収集されたパターン対の数

上記の方法では，861 の和語動詞に対して，5,243 文（和文 7.5 万字，英文 4 万語）の対訳例文が得られた。これらの対訳用例を使用したパターン対作成作業では 1,399 パターン対が新規に作成され，既存のパターン対のうち 414 件が修正された。

(3) 追加拡充の程度

IPAL 動詞辞書は，日本語動詞の語義分類に基づいて用例が作成されている。したがって，日英翻訳用のパターン対の観点から見ると，日本語動詞の語義とパターン対との対応関係（1 語義が 1 パターンに対応するか）が問題となる。そこで，日本語の語義の多い 4 動詞について，語義と

26 日本語の動詞において語彙体系上ならびに使用頻度上重要であると考えられる基本的な 和語動詞 861 語（ひらがな表記した場合で，漢字表記では 1,301 語に相当する）について，意味および統語的特徴に基づいて下位区分し，それを 1 つの単位として，意味，形態，統語，文法カテゴリ，慣用表現などに関わる情報が詳細に記述されている。また，各下位区分ごとに 1～3 文の用例が付されている。

パターン対の対応関係を調査した。その結果を表 **4.1** に示す。この表から，日本語用言の語義とパターン対が 1 対 1 に対応するものは 4 割にとどまり，両者は必ずしも対応しないことが分かった。このことは，日英機械翻訳から見れば，IPAL 辞書の語義分類は，英語に訳出するうえで，必ずしも適切ではないことを意味している。すなわち，日英翻訳では，日本語と英語の意味的対応関係に即して，日本語の語義分類をする必要のあることが分かる。

表 **4.1** **IPAL** 語義と文型の対応

分類 / 動詞	「語義」対「パターン」の関係					合計
	1 対 1	1 対 *n*	*m* 対 1	*m* 対 *n*	保留	
あがる	8	5	1	3	1	18
あげる	14	2	1	1	3	21
だす	8	9	5	4	1	27
でる	13	3	10	4	2	32
合計	43 43.9%	19 19.4%	17 17.3%	12 12.2%	7 7.1%	98 100%

4.3.3 人の知識を内省する方法

(1) パターン用例収集の方法

前節までの結果から，人間用の和英辞書，日本語辞書の双方から用例を収集しても十分なパターン対が作成できないことが分かった。日本語例文とパターン対の関係を観察すれば，同じ動詞を使用していても，動詞の使われ方のニュアンスが異なるときに新たな英語パターンが必要となる場合が多いことに気がつく。そこで，第 3 の方法として，英語の理解できる日本人が辞書等を参考にしながら自分の知識を引き出し，日本語としてニュアンスの異なる用法を可能な限り列挙するという方法で用例の収集を試みる。

列挙する用例は，作業にかける時間にもよるが，ある程度の時間以上考えても用例が思い浮かばなくなるまで抽出することとした[27]。用例数としては，いくつかの動詞について思考実験した結果に従い，IPAL 動詞辞書の語義数の 2～3 倍を目標とした。また，これらの日本語用例に対する英訳は翻訳専門家に依頼し，対訳用例集を作成することとした。

(2) 収集されたパターン対の数

上記の方法による作業結果では，約 1.5 人年の作業により，861 動詞[28]に対し用例 10,500 文

27 後に述べるように，経験によれば，作業開始当初は「出る」「上がる」「掛ける」など，語彙数の多い動詞の場合，1 動詞の用例を書き出すのに 1 人日程度かかった。しかし，慣れてくるにつれて速くなったこと，通常の和語系動詞はそんなに語義がないことにより，1 日平均で 3 動詞前後の用例が抽出できるようになった。

28 見出し語は仮名表記のめ，漢字仮名混じりの表記では語数は増大する。実際に収集した漢字仮名混じりの語

(和文 13 万字, 英文 6.8 万語)が収集された[29]。

また, 収集した用例から, 語義数の多い動詞と少ない動詞が混合するように 36 動詞(1,100 用例文)を選び, パターン対の抽出を試行したところ, 新たに 300 パターンが抽出された。第 1 ならびに第 2 の方法で得られなかったパターン対が 1 動詞当たり平均 10 パターン見つかったことになる。

4.4 収集された用例とパターン対の数の比較推定

4.4.1 収集された用例の数とパターン対の数の比較

前述の 36 の和語動詞について, 3 章で述べた 3 つの方法によって得られた用例数とパターン対の数を比較して表 **4.2** に示す。参考のため, この表のほぼ中間位置にある動詞「上がる」について, 3 種の方法で得られた動詞用例とそれらの用例から得られたパターン対の関係を付表 **4.1**～**4.3** に示す。

表 **4.2** から, 各手法で得られたパターン対の数の関係を示すと図 **4.4** のとおりとなる。これらの図表より, 以下のことが分かる。

① 第 1 の方法に加えて第 2 の方法を実施すれば, 第 1 で得られるパターン対の数の約 2 倍のパターン対が収集できる。

② 第 1, 第 2 の方法に加えて第 3 の方法を実施すれば, 第 1, 第 2 で得られるパターン対の数のさらに 2 倍以上のパターン対が収集できる。

これらの結果から, 和語動詞について見れば, 和英辞書がら収集されるパターン対の約 4 倍が人間の知識の内省によって得られることになる。

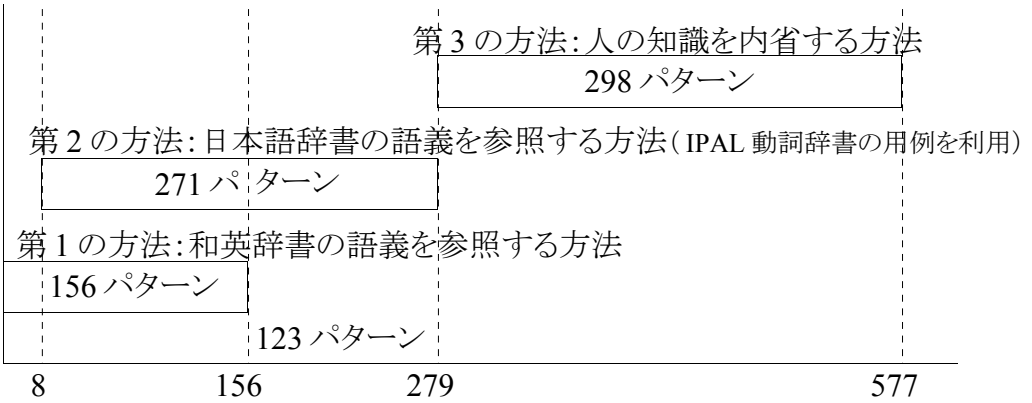


図 **4.4** **3** 種の方法で得られたパターン対の種類の関係

数は約 1,100 語である。

29 人件費で見ると, 和文用例作成のコストとその英語への翻訳コストはほぼ同じである。

表 **4.2** 収集されたパターン対の数の比較(和語動詞の例)

	収集方法	第 1	第 2			第 3		一般 P 数	[参考]
IPAL 表記	漢字表記	P 数	語義数	例文数	追加 P	用例数	新規 P	数合計	慣用 P 数
でる	出る	22	32	49	5	145	38	65	18
だす	出す	16	27	53	15	95	22	53	21
あける	空ける 明ける 開ける	4 4 3	11	17	1 0 1	14 9 9	5 2 2	10 6 6	0 1 1
たつ	立つ 発つ 建つ 経つ	5 2 1 2	13	24	4 0 0 0	75 6 5 3	30 1 0 0	39 3 1 2	11 0 0 0
あく	空く 開く	4 5	10	12	4 2	13 12	4 4	12 11	1 0
たてる	立てる 建てる	8 1	9 1	17	7 0	69 5	29 0	44 1	7 0
あげる	上げる	8	21	31	13	98	16	37	14
おちる	落ちる	8	11	21	7	53	23	38	1
たつ	断つ 絶つ	4 4	1 3		0 0	6 6	1 2	5 6	0 0
あがる	上がる	7	18	31	16	90	16	39	12
はいる	入る	7	23	34	11	105	31	49	5
おとす	落とす	6	14	19	5	53	15	26	3
くずす	崩す	6	4	4	2	8	2	10	0
いれる	入れる	5	19	30	12	113	28	45	10
くずれる	崩れる	5	4	6	2	13	4	11	0
きめる	決める	3	14	20	4	28	5	12	0
さける	避ける	3	6	11	0	9	2	5	0
きまる	決まる	3	8	17	2	32	10	15	2
うめる	埋める (うめる) (うずめる)	3	 4 4	 5 4	1	9	0	4	1
さく	割く 裂く	2 1	5	7	0 5	4 3	2 0	4 6	0 0
うまる	埋まる (うまる) (うずまる)	2	 5 3	 6 4	2	5	1	5	0
さける	裂ける	1	1	3	2	4	1	4	0
さく	咲く	1	1	1	0	3	2	3	0
	合計	156	271	426	123	1102	298	577	108

4.4.2 必要なパターン数と用例数の見積もり

(1) 和語動詞の場合

第 3 の方法で得られるパターン対の網羅性を調べるために，アナリストを代えて用例作成を行い，その用例から得られたパターン対を比較した。その結果，各アナリストの作成した用例から得られたパターン対はほぼ一致することが分かった[30]。したがって，前章で取り上げた個々の動詞のパターン対の数は，ほぼ，それぞれの動詞に必要なパターン対の数とみなせる。この結果に基づき，和語動詞に対して日英機械翻訳でどれだけの数のパターン対が必要とされるかを予測する。

まず，第 1 の方法で得られたパターン対の数を図 **4.5** の実線で示す。次に，前節で取り上げた動詞に対して第 2，第 3 の方法で得られたパターン対の数をプロットし，それらの点をなめらかに結べば，それぞれ波線，一点鎖線の結果が得られる。

この図から，和語動詞に対して必要となるパターン対の数はおおよそ 9,000 件と推定される。

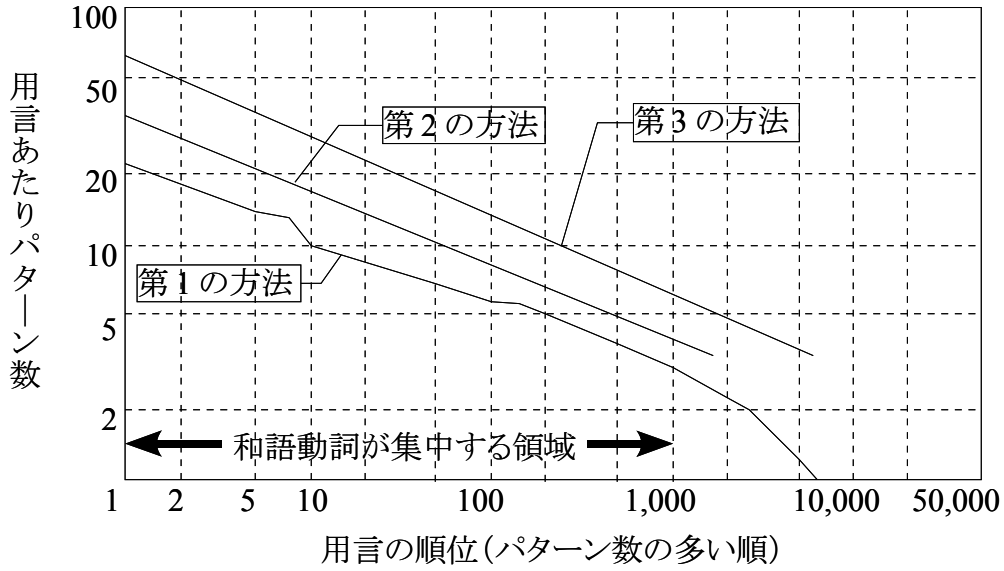


図 **4.5**　日本語用言に対するパターン対の数の分布

(2) 最終的な規模予測

日英機械翻訳システムにおいて，パターン対に整理することが適切と見られる述語としては，和語動詞のほかに漢語動詞，形容詞系の述語等がある。また，前章までは一般パターンについて述べたが，用言性の慣用表現もパターン対とすることが適切と考えられる。

これらのうち，形容詞系の述語は和語動詞と同様の性質を持つため，本章と同様の方法が適切

30 意味属性の指定等では揺らぎがあり必ずしも一致しないが，必要なパターンの種類ではほぼ一致する。例えば，20～30 パターンを持つ動詞の場合，2 人のアナリストが独立に作成した用例から得られたパターン対のうち，一方が作成したが他方が作成しなかったパターン対は 1～2 件程度（再現率 90%以上）である。パターン対数の少ない用言の場合の再現率はさらに高い。

と考えられる。慣用パターンも同様である。漢語動詞は，通常，1 単語当たりのパターン対数はほぼ 1～2 件であるため，用例が得られれば比較的容易に収集可能である[31]。

そこで，これらの語を含む用言全体として必要と見られるパターン対の数を推定すると表 **4.3** を得る。表 **4.3** では，推定されたパターン対の数に対して，本章の方法でどれだけ収集できる見込みかについても示す。

この表から，日英機械翻訳では，一般パターン，慣用パターンを含め，約 25,000 のパターン対が必要と推定される。

表 **4.3**　日英機械翻訳で必要なパタン対の数とその収集に必要な用例の数の見積もり

（* 漢字表記による異なり数）

比較項目 / パターンの種類		必要量見込み		第 1 の方法		第 2 の方法（追加見込み）			第 3 の方法（追加見込み）		
		用言数	パターン対数	用言数	パターン対数	*対象用言数	用例数	パターン見込	対象用言数	見込み用例数	パターン見込
一般パターン	和語動詞	1,500	9,000	1,500	4,000	1,000	5,200	1,500	1,200	15,000	3,500
	サ変動詞	6,500	8,000	3,000	4,000	(50)	(141)	(9)	4,000	8,000	4,000
	形容詞系	2,000	3,000	1,100	2,000	200	2,400	500	500	2,000	500
	小計	10,000	20,000	5,600	10,000	1,200	7,600	2,000	5,700	25,000	8,000
慣用パターン		---	5,000	---	3,000	---	---	---	---	不明	不明
合計		10,000	25,000	5,600	13,000	1,200	7,600	2,000	5,700	25,000	8,000

4.5 用例の収集方法

4.5.1 用例の網羅的収集

一般パターン対の網羅的の向上を考えたとき，抽象化した状態の文型を収集するのは容易ではないため，まず様々な用法を例文という形で内省により網羅し，例文を抽象化する 2 段階で考える。例文作成の対象とする用言の選択として，現代語の用言として相応しいかどうかは個別に判定するが，1 つの辞書を選択して大まかな基準として利用する方針とした。また，生成された例文が自然な表現かどうかが問題となる場合はあった。これについては，同じ作業者がある程度時間が経ってから見直すか，別の作業者と相互検査することにより排除に務めた。一連の作業経験を踏まえ，次の条件を設定した[2]。

① 現代国語例解辞典[68,69]所収の用言性の語を対象とし，語釈や例文を参照し，または類推することにより例文を作成する。

＜備考＞自然な例文を作成できる語を対象とした。収録語のうち現代語として不適当と思わ

31 また，パターン対間の相互作用の心配は少ないため，パターン対当たりの作成効率は和語動詞系の用言の場合に比べてはるかに良い。

れる語を除外した。例文の作成を進めながら，例文作成が困難なものを，例文作成者同士の協議により除外した。

② 例文作成者の主観で，用言のニュアンスが異なると感じられるものを広く例文化し，可能な限り「一般的で単純な名詞を格要素とする単文」とする。

＜備考＞日本語表現の作成作業として位置付ける。すなわち，対応する英訳が異なるところまでは要求しない。結果的に訳語が同じになっても許容する。

③ 用言が終止形で終わる例文だけでなく，連用形や連体形の用法のニュアンスの違いにも留意して例文を作成する。

＜備考＞連用形の副詞用法や連体形の限定用法には慣用的なものがあり，それらの収集も対象とした。

④ 用言 1 語当たりの 2 例文を最低目標とする。ただし，ある程度考えても例文が思いつかなくなるまで作成を行なう。

＜備考＞これまでの経験では，n 文の作成時間を t として，おおよそ t は n^2 に比例する。10～15 分考えても新たな用法が思いつかなければその用言に対する作業を打ち切ることにした。

⑤ 収集された例文に対して，可能な限り原文に忠実で，かつ，英語として十分通用するように，翻訳家に英訳してもらう。（最低限度の意訳は許容する）

＜備考＞経験を踏まえ，英語を母語とする翻訳家と日本語を母語とする翻訳家の共同作業に委ねた。

4.5.2 多様な表現の収集

最も直接的な動機は，1 つの日本語表現に対する複数の英訳を得ることである。これは見方を変えれば英語表現を換言しているとも云える。一方，ある英語表現が必ず特定の日本語表現から生成されなければならないわけでもない。そこで，日本語の換言と英語の換言を並行して実施することにした。

換言事例の収集という課題は，元来は同一言語内で何らかの観点で同義の表現を収集するべきかもしれない。しかし，例文が提示されるとそれに眩惑されて他の表現がなかなか思いつかない場合も少なくないし，観点の種類をあらかじめ列挙しておくことも難しい。そこでここでは日英の対訳対の存在を前提として，その日英の文対を制約として利用しながら，いわば多様な翻訳例文の作成として，換言事例を収集することとした[3]。

ここでいう換言は，例えば英作文の際，和英辞書に載っていない語や表現に出会ったとき，日本語の別の類義表現を生成し，和英辞書を引き直すことを模したものである。したがって，翻訳対象言語に精通していない単言語話者にも作業可能であると考えられる。しかし実際問題として，考えついた別の表現が和英辞書に未集録であるという状態が連続して発生すると，同義性の制約が徐々に甘くなっていく，すなわち意味のずれが拡大していく恐れがある。そこで，今回は試行ということもあり，網羅的な用例収集の作業担当者とその例文の翻訳担当者に依頼した。これは網

羅性の確保の際に微妙な日英対応の判定が容易でなかった経験に基づく。また，以下の条件設定は今回の問題点の分析を踏まえて改善していきたいと考えている。

① 前節で述べた日本語の用言に対する日英の対訳例文対を対象とする。
② 日本語の換言は，英文に多様な和訳をつけるつもりで行なう。逆に，英語の換言は，日本語例文に多様な英訳をつけるつもりで行なう。
③ 原則として，特殊な場面設定を必要としない中立的な表現を作成する。

4.5.3 収集の経過と結果

用言ごとの様々な用法が例文として収録されているIPAL辞書[66,70]に着目し，ニュアンスの異なる用言の用法を例文として追加した。次いで，国語辞書に基づいて用言の網羅性を高めることとし，用言選択の基準を現代国語例解辞典[68,69]に置いた。IPAL辞書に収録されていない用言を対象にして例文作成を継続中（現在はサ変動詞）である。また，途中から換言作業も並行して進めている。

表 **4.4** に収集情況を示す。ただし「和語動詞/IPAL」は和語動詞のうちIPAL動詞辞書に収録されている語，「和語動詞/他」それ以外の語を対象としたことを示す。なお，備考に作業順と作業内容を示す。各項目は1～3人年の作業量であった。ただし，一部並行して実施したものもある。サ変動詞に対する換言は，他との比較では比較的容易であると言える。付表 **4-A**～**4-E** に例文を示す。

表 **4.4** 用言の種類と例文数

用言種別	該当用言数	作成例文数	換言例文数		換言文なし		備　考（作業内容）
			日本文	英文	日本語	英語	
和語動詞/IPAL	849	16,713	7,043	4,096	12,020	13,748	追加後に換言済み
和語動詞/他	936	1,883	0	0	1,883	1,883	収集済み（未換言）
複合和語動詞	2,101	3,701	1,212	480	2,487	3,220	収集後に換言済み
イ型形容詞/IPAL	136	2,156	530	219	1,626	1,937	追加後に換言済み
イ型形容詞/他	522	830	1,561	1,584	1	0	収集後に換言済み
ナ型形容詞	1,296	2,356	621	440	1,735	1,915	収集後に換言済み
（サ変動詞=途中）	(885)	(1,550)	(4,448)	(4,245)	(6)	(3)	収集・換言を並行中
合計	5,840	27,639	10,967	6,819	17,869	20,820	（注）サ変動詞を除く

4.6 結言

日英機械翻訳において，用言（動詞，形容詞）の意味を訳し分けるのに必要な結合価パターン対の数とそれを収集する手段について検討した。

具体的には，単語当たりの語義が多いためパターン対作成が最も困難な和語動詞の場合を取り上げ，(1)和英辞書から収集する方法，(2)日本語動詞の語義対応の用例を使用する方法，(3)それらを参考に，人の知識に基づいて用例を作成して使用する方法の3種のパターン対の収集方法を比較した。その結果，主要な約1,000の和語動詞を意味によって訳し分けるには7,500件の結合価パターンが必要であることが分かった。これに対して，従来の和英辞書から収集できるパターン対の数は約1/4，和英辞書と日本語辞書の語義分類知識を使用する場合は約1/2であること，必要なパターン対を網羅的に収集するには，作業工数の面でも，和英辞書と日本語辞書の語義を参考に人の知識を内省して用例を作成する方法が適していることなどが分かった。

また，上記の結果から推定すると，漢語動詞，形容詞系の述語，用言性慣用表現などを含むパターン対全体では約25,000パターンが必要なこと，それらのパターンも辞書等を参考に人の知識を内省する方法で抽出された用例から比較的容易に収集できる見込みであることが分かった。

なお，現在，第1の方法で得られたパターン対を拡充するため，第2，第3の方法を並行して実施中であり，和語動詞，漢語動詞，形容詞系述語に対してそれぞれ，5,500件，4,000件，2,000件（合計11,500件）のパターン対を収集済みである。また，慣用表現では約3,000のパターンが収集されている。今後は，残されたパターン対（一般パターン約8,500件，慣用パターン約2,000件）を整備していく予定である。

付表 **4.1** 最終的に得られたパターン対(「上がる」の場合)

パターン番号	方法	パターン対(意味属性などの条件やパターンの構造は省略)	
		日本語パターン	英語パターン
P01	①	AがBからCに上がる	A rise from B to C
P02	①	Aが上がる	A go up
P03	③	AがBをCに上がる	A go up B to C
P04	①	AはBが上がる	A produce good B
P05	①	AがBで上がる	A get nervous at B
P06	①	AがB[数量]CからDに上がる	A be raised by B from C to D
P07	①	AがBに上がる	A appear as B
P08	①	Aが上がる	A be dead
P09	②	AがBに上がる	A splash over B
P10	②	AがBに上がる	A appear on B
P11	③	Aが上がる	A be raised
P12	②	Aが上がる	A stop
P13	②	AがBからCに上がる	A improve from B to C
P14	②	AがBを上がる	A would like some B
P15	②	AがBに上がる	A enter B
P16	②	Aが上がる	A arise
P17	②	AがBで上がる	A be completed in B
P18	②	AはBからCにD[地位]が上がる	A be promoted from B to C
P19	②	AがBで上がる	B be enough for A
P20	②	AがBを上がる	A fly into B
P21	②	AがBに上がる	A be landed on B
P22	②	Aが上がる	A be produced
P23	②	Aが上がる	A be arrested
P24	③	Aが上がる	A be sluggish
P25	②	AがBで上がる	A die as a result of B
P26	③	Aが上がる	A increase
P27	③	AはB[男ぶり]が上がる	A improve in A's look
P28	③	AはB[気勢]が上がる	A be in high spirits
P29	③	AがBから上がる	A be collected from B
P30	③	AがBに上がる	A go to B
P31	③	AがB[時代]をCに上がる	A go back to C
P32	③	Aが上がる	A end
P33	③	Aが上がる	A rise
P34	③	AはBが上がる	A go out of A's B
P35	②	AがBから上がる	A come out of B
P36	③	AがBに上がる	A be on B
P37	③	Aが上がる	A be found

(注)パターン番号はパターン選択をする際の優先順位を示す。

付表 **4.2** 第 **2** の方法によるパターン対作成（**IPAL** 用例「あがる」）

No	IPAL語義	日本文用例	英語訳	既	新
1	生き物が上方に移動する。	一行は階段を一階から五階に上がった。	The party went up the stairs from the 1st floor to the 5th floor.	P02	
		彼は坂道を一気に上がる。	He climbs slopes without stopping.	P02	
2	物が上方に移動する。	水銀柱が三十度に上がった。	The mercury in the thermometer rose to 30 degrees.	P01	
		花火が夜空を空中高く上がっていく。	Fireworks are flying high into the sky.		P20
3	ある事柄の程度が高くなる。	会社は生産が上がった。	The company increased production.		P13
		勉強の能率が上がった。	Study efficiency has improved.		P13
4	今までよりも上の段階になる。	国鉄は初乗り運賃が120円から140円に上がった。	JNR raised its basic fares from 120 yen to 140 yen.	P06	
		アパートの家賃が1万円上がった。	The apartment rent increased by 10,000 yen a month.	P06	
5	今までよりも上の段階になる。	彼は係長から課長へ地位が上がった。	He has been promoted to section chief from chief clerk.		P18
		娘の算数の成績が4から5に上がった。	My daughter's mathematics mark improved from four to five.		P13
6	（上の）学校に進む。	娘は今年小学校に上がる。	My daughter will enter elementary school this year.		P15
7	（ある目的のために）ある場所に入る。	落語家が高座に上がる。	The storyteller appears on the stage.		P10
		友人は歌手として舞台に上がった。	My friend appeared on the stage as a singer.		P10
8	水の中から出る。	子供が風呂から上がった。	The child stepped out of the tub.		P35
		海亀が海から陸に上がる。	Sea turtles land on shore.		P21
9	ある現象が発生する。	会場に歓声が上がった。	A shout of joy arose in the hall.		P16
		辺りに水しぶきが上がった。	A spray of water splashed around.		P09
10	好ましい結果が得られる。	こうすれば利益が上がる。	If you do so, a profit will be obtained.	P04	
11	候補として名前が出る。	彼は次期社長の候補に名前が上がる。	He is running as a candidate for president.	P07	
		Aチームの名前が代表候補に上がった。	Team A was nominated as a representative team.	P07	
12	今まで分からなかったものが明らかになる。	証拠が挙がった。（注1）	Evidence was deduced.		P22
		犯人が挙がる。（注1）	The suspect was arrested.		P23
13	何かが完成・完了する。	夕立が上がった。	The shower has stopped.		P12
		原稿が上がった。	The manuscript has been prepared.		P17
14	ある数量で済む。	会費が4000円で上がった。	Four thousand yen was enough for the membership fee.		P19
		設置は二時間で上がる。	It takes two hours to install.		P17
15	活動機能が停止状態になる。	赤潮で魚が上がった。	Fish died as a result of red tide.		P25
		バッテリーが上がる。	The battery is dead.		P08
16	興奮状態に陥る。	私は入試で上がってしまった。	I got nervous at the entrance examination.	P05	
17	「食べる，飲む，喫う」の尊敬表現。	ビールを上がりませんか。	Would you like some beer?		P14
18	「行く，訪ねる」の謙譲表現。	私がお届けに上がります。	I will deliver it.	（注2）	

（注 1）「挙がる」は「上がる」の表記の揺れとして扱う。（注 2）日本語のニ格の条件が複雑なため登録を保留している。

付表 **4.3**　第 **3** の方法によるパターン対作成(「あがる」の全 **90** 用例の一部)

No	日本文用例	英語訳	既存	新規
1	時代を上がる。	Move back in time.		P31
2	梅雨が上がった。	The rainy season ended.		P32
3	物価が上がった。	Prices went up.	P02	
4	歓声が上がった。	(The crowd) cheered.		慣用
5	遺体が上がった。	The body was found.		P37
6	悲鳴が上がった。	(The girl) screamed.		慣用
7	犯人が上がった。	The criminal was found.		P37
8	娘が屋敷に上がる。	The girl goes up to the mansion.	P02	
9	7 時に幕が上がる。	The curtain rises at 7:00.		P33
10	年貢は領地から上がる。	Land taxes are procured from the territories.	P22	
11	ダムの水位が上がった。	The water level of the dam rose.		P33
12	列車のスピードが上がった。	The train's speed increased.		P26

付表 **4-A** 和語動詞の対訳換言例文(一部)

J0　彼の企画が当たった。 J1　彼の企画が成功した。	E0　His plan was a success.
J0　彼はその漢字を辞書に当たった。 J1　彼はその漢字を辞書で調べた。	E0　He looked up that character in the dictionary.
J0　私は彼の行き先について友人たちに当たった。 J1　私は彼の行き先について友人たちに聞いた。	E0　I asked his friends about his destination. E1　I questioned his friends about his destination.
J0　彼は暑さにあたった。 J1　彼は暑さ負けした。	E0　He was affected by the heat.
J0　私の予想が当たった。	E0　My prediction was right.
J0　彼はふぐにあたった。	E0　He was poisoned by eating blowfish.

付表 **4-B** 複合和語動詞の対訳換言例文(一部)

J0 J1	競技場は大勢の観客で膨れ上がった。 競技場は大勢の観客で身動きできなかった。	E0	The athletic field was swamped with spectators.
J0	蜂にさされたあとが膨れ上がった。	E0	The place where I was stung by the bee has swollen up.
J0 J1	この都市の人口は 10 年前の 2 倍に膨れ上がった。 この都市の人口は 10 年前の 2 倍だ。	E0 E1	The population of this city is double what it was 10 years ago. The population of this city has doubled in the last 10 years.

付表 **4-C** イ型形容詞の対訳換言例文（一部）

J0	彼の態度は好ましい。	E0	His attitude is favorable.
J0	彼は我が社には好ましくない人物だ。	E0	He is not the kind of person we want in our company.
J0	ディナーには正装が好ましい。 ディナーには正装が望ましい。	E0	Formal attire is desirable for dinner.
J0 J1	ジャガイモは常温での保存が好ましい。 ジャガイモは常温での保存が最もよい。	E0 E1	It is best to keep potatoes at room temperature. Potatoes should be kept at room temperature.

付表 **4-D** ナ型形容詞の対訳換言例文（一部）

J0	私は今の地位に満足だ。	E0	I am satisfied with my present position.
J0 J1	私は昨日から満足な食事をしていない。 私は昨日からまともな食事をしていない。	E0 E1	I have not had a proper meal since yesterday. I have not eaten a proper meal since yesterday.
J0 J1	彼はアルファベットも満足に書けない。 彼はアルファベットもろくに書けない。	E0	He cannot even properly write the alphabet.

付表 **4-E** サ変動詞の対訳換言例文（一部）

J0 J1 J2	彼らの攻撃は相手チームを圧倒した。（ｽﾎﾟｰﾂ） 彼らの攻撃は相手チームを圧した。 彼らの攻撃は相手チームをねじ伏せた。	E0 E1 E2	Their attack overwhelmed the opposing team. Their attack overpowered the opposing team. Their attack swamped the opposing team.
J0 J1 J2	私はナイアガラ瀑布の壮大さに圧倒された。 私はナイアガラ瀑布の壮大さに威圧された。 私はナイアガラ瀑布の壮大さに気圧された。	E0 E1 E2	I was overwhelmed by the scale of Niagara Falls. I was thunderstruck by the magnificence of Niagara Falls. I was awed by the scale of Niagara Falls.
J0 J1	シートベルトが腹部を圧迫する。 シートベルトが腹部を押さえつける。	E0 E1 E2	The seatbelt is pressing into my stomach. The seatbelt is pressuring my stomach. The seatbelt is digging into my stomach.

第 5 章　連鎖型および離散型共起表現の自動抽出

5.1 緒言

最近，自然言語処理において，大量のコーパスや用例の重要性が指摘され，それを分析する技術の必要性が増大している。例えば，機械翻訳では，単語単位の直訳ではうまく訳せないフレーズを集め，フレーズ単位に翻訳する方法や，一定の構造を持つ表現を対訳パターン化し，パターン辞書によって原言語を目的言語に対応づける方法などが考えられている。これらの方法を実現するには，現実に使用されている言語データの中から，使用頻度の高いフレーズや表現のパターンを抽出することが必要である。

しかし，膨大な言語データを対象とするとき，任意の長さで，出現頻度の高い表現文字列を漏れなく自動的に発見して，抽出することは，計算量の点で困難であった。そのため，従来，自然言語としての特徴に着目する方法，抽出する文字列の性質に着目する方法など，目的に合致する文字列を限定的に抽出する方法が考えられてきた。例えば，前者の方法としては，言語データから結びつきの強い単語を取り出す観点から，2 単語の結びつきの強度に着目した方法[71]，単語間の距離に着目した方法[72]，結合単語数と出現回数を考慮した方法[73,74]などが提案されている。後者の方法としては，抽出する単語や文字の連鎖の数を制限したり，短い連鎖で出現頻度の高いものに着目して，限定された文字列（単語列）の範囲で連鎖数を増やして集計する方法[75]などが考えられていた。

これに対して，最近，大量の言語データを対象に，任意の n に対する n-gram 統計を高速に実行する方法が提案され[15]，言語データ内にある任意の長さの文字列（一般には記号列）を自動的に抽出し，その出現回数をカウントすることが可能となった。この結果を用いれば，原文中に使用された文字列を，その長さ（文字数）の順かつ出現頻度の高い順に集計することができる。しかし，この方法では，抽出する文字列間の相互関係が無視されているため，すでに抽出された文字列の部分文字列が重複して抽出される。したがって，抽出された文字列を言語表現として見た場合，文法的，意味的にまとまりのない断片的な文字列が多数を占める。これを意味のある文字列に絞り込む方法として，同一の論文[15]では，抽出された文字列とその出現回数を相互に組み合わせる方法の可能性を示している。その後，この n-gram 統計データとして得られた文字列から意味のある表現を取り出す方法として，抽出した文字列のエントロピー基準を用いる方法[76]が提案されている。また，n-gram 統計を応用したものに，助詞的定型表現の抽出の例[77]があるが，この方法では，あらかじめ，抽出する文字列を構成する字種の組を限定することで n-gram の計算量の問題を回避し，その後，抽出された文字列を種々のヒューリスティックスを用いて絞り込んでいる。

次に，離れた位置に共起する表現の組の抽出を見ると，複数の文字列を組み合わせて，原文中での共起を調べることが必要である。n-gram 統計では，膨大な量の文字列が抽出されるため，

抽出された文字列すべてを組み合わせて原文をサーチするのは物理的に困難であった。連鎖型，離散型を特に区別せず，1 文中に共起表現が占める割合の多い文を定型的な文として抽出する試み[78,79]もあるが，大量の言語データの中から，出現頻度の高い文字列の組を，漏れなく自動的に発見し集計するのに効果的な方法は知られていない。

ところで，大量の言語データを対象とするとき，共起表現抽出の問題は，第 1 に，計算量（ファイル量）増大による計算可否の問題であり，第 2 に，得られた大量の結果から必要な表現を選択する問題である。特に，離散型共起の場合，計算量は，それを構成する表現要素の数に対して幾何級数的に増加することが問題となる。計算量を削減する方法を考える際は，共起表現抽出の目的から考えて，必要な共起表現を漏らしてしまうような絞り込みは望ましくない。

連鎖型の文字列抽出の場合は，n-gram 統計の方法によって，すでに，第 1 の問題は解決されているが，表現の単位とみなせない（単語の断片を含む）断片的な文字列が多数抽出される。このため，離散型共起の場合，計算量が増大し，計算不可能となることが問題となる。断片的な文字列の抽出が抑制され，計算量が可能な範囲に収まれば，離散型共起においても，第 1 の問題は解決する。また，第 2 の問題については，最終的には，使用目的ごとに人手で判断せざるを得ないから，出力される文字列の量（種類）が，人手作業に支障のない範囲（数千種，最大数万種以下）になれば，第 2 の問題も当面解決したと言える。

以上の観点から，本節では，連鎖型共起表現抽出において断片的な文字列抽出を抑制する方法として，言語データの中から，最長一致の文字列抽出（ある文字列が抽出されたとき，その文字列に含まれる部分文字列は抽出しない）を条件とし，任意の長さ以上，任意の使用頻度以上の共起表現を，漏れなく，自動的に抽出し，集計する方法を提案する。次に，その結果を使用して，複数の要素が離れた位置に共起する離散型共起表現を自動的に抽出し集計する方法を示す。また，提案した手法の動作確認のための適用例として，日本語新聞記事データからの連鎖型，離散型共起表現の抽出結果を示す。

5.2 従来の方法とその問題点

(1) 文字列抽出の条件

自然言語の文中で共起する表現[32]としては，連語やフレーズのように連続した文字列を構成するもの（連鎖型共起表現と呼ぶ）と，係り結び，呼応関係，特定の動詞と特定の名詞の組などのように，2 種類以上の文字列が，文中の離れた位置に現れるもの（離散型共起表現と呼ぶ）がある。離散型共起表現は，連鎖型共起表現の文字列が文中で共起したものと考えることができるから，まず，前者の文字列を考える。

さて，連語やフレーズのような連続した文字列を漏れなく発見すること，また，文法的，意味的に見て，表現の単位をなさないような断片的な文字列の抽出を最小限に押さえることを狙って，以下の条件で文字列を抽出することとする。

第 1 の条件：任意の長さ以上の文字列を抽出する。

第 2 の条件：任意の出現頻度以上の文字列を抽出する。

第 3 の条件：最長一致の原則で文字列を抽出する。

このうち第 3 の条件は，原文中のある場所からある文字列が一度抽出された後は，その文字列内に含まれる部分文字列は抽出の対象としないことを意味する。ただし，その部分文字列が別の場所に現れたときは抽出される。例えば，図 **5.1** の場合，7gram の文字列 α が抽出されたとすると，それ以降の 6gram 以下の文字列の抽出では，α 部分の部分文字列である β や γ は対象外とする。ただし，α が抽出された場所以外の位置に現れた「DE」，「GHI」は当然，抽出の対象となる。また，文字列 δ は，α の部分文字列でないので，抽出の対象とする。

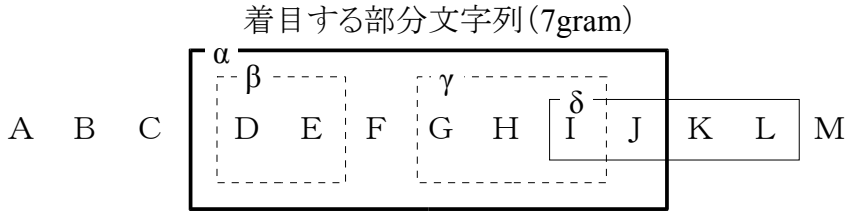


図 **5.1** 抽出対象文字列の例

(2) 表現抽出における最長一致の原則の意義

一般に言語表現は，大小の表現が幾重にもネストして構成される。共起表現の抽出では，表現の単位や長さをあらかじめ指定しなくても，このような表現の中から，繰り返し使用される表現の単位を自動的に発見し，抽出できることが望まれる。そこで，すべての文字列を網羅的に抽出すれば，そのような表現は抽出されるが，一度抽出された文字列の中からも部分文字列が重複して抽出されるため，多くの断片的な文字列が含まれることが問題となる。

32 本章では，文法的，意味的に表現の単位とみなせる文字列を意識して「表現」と呼ぶ。

ところで，言語の共起表現は，複数の単語が共起した表現だと考えると，共起表現の文字列の境界は，同時に単語境界ともなっている。一方，可能な限り長い単位で文字列を抽出すれば，その文字列の境界は単語境界に一致する可能性が高いから，断片的な文字列ではなく共起表現である可能性が高くなる。すなわち，断片的文字列の抽出が抑制されると期待される。以上から(1)では，第 3 の条件を設けた。

ここで，第 3 の条件で抽出が抑制される文字列について考える。抑制される文字列には，より大きな文字列の部分としてしか使用されないため，一度も抽出されないものと，他の部分からは独立性のある表現として何回か抽出されるが，ある文字列の部分文字列として使用された部分でカウントが抑制されろものがある。共起表現の網羅性の観点から見れば，このうち，前者の抽出漏れが問題で，その中に，表現とみなせる文字列が含まれるかどうかが大切である。

しかし，ある表現がより大きな文字列の中に埋もれた部分的な表現であっても，独立性が高く，繰り返して使用されるような表現であれば，ある文字列の部分文字列としてだけでなく，それ自身が最長の単位であるような文字列として繰り返し出現することが期待できる[33]。以上から，第 3 の条件があっても，繰り返し使用される共起表現（の種類）は，網羅的に抽出されるものと期待できる。

(3) 長尾・森の方法とその問題点

任意の n に対する n-gram を効率的に抽出して集計する方法として，すでに，長尾・森の方法[15]が提案されている。この方法を要約すると以下のとおりである。

［長尾・森の方法］

集計対象とする言語データ全体の文字数を N とする。

手順 **1:**「原文番地ファイルの作成」

N 個のレコードからなるファイル（原文番地ファイル）を用意し，各レコードに，0 から順に N-1 の値（原文番地）をいれる。原文番地は，言語データ上，その値で示される文字番号から始まり，末尾（N-1 番目の文字）で終わる部分文字列（以下，文字列単語と呼ぶ）へのポインタの意味を持つ。

手順 **2:**「汎用ソートファイルの作成」

原文番地ファイルの各レコードを，対応する文字列単語の文字コード順に，ソートしたファイル（汎用ソートファイル）をつくる。

手順 **3:**「一致文字数のカウント」

汎用ソートファイルの各レコードの示す文字列単語を，その直後のレコードの文字列単語と先頭文字から比較し，一致した文字数（一致文字数）を書き込む。

手順 **4:**「文字列の抽出とカウント」

一致文字数をレコード順に調べ，部分文字列の種類とその出現回数を編集する。

この方法により，任意の回数以上出現した文字列を長さ（文字数）ごとに，かつ，出現回数の大

33 ある表現の部分としてしか使用されないような部分的な表現は抽出されないが，そのような表現は，元々，それを含むより大きな表現の一部にすぎないと考えられるから，改めて取り出すことはしない。

きい順に得ることができるため，目標とする第 1，第 2 の条件は満足されるが，目標とする第 3 の条件は満足されない。

第 3 の条件を満たすようにするため，抽出された文字列の出現回数に対して，より長い文字列に含まれていた部分文字列の出現回数を差し引くなど，次数の異なる複数の *n*-gram 集計表を組み合わせて計算する方法が考えられるが，集計表が生成された時点では，抽出された文字列の原文中での相互関係の情報が失われているため，計算は不可能である[34]。

5.3 連鎖型共起表現の抽出法

5.3.1 重複する文字列の扱い

汎用ソートファイルに戻って，言語データの中で，一度抽出した文字列の部分は別の文字列として改めて抽出したり，カウントしたりしない方法を考える。以下では，汎用ソートファイルから，一致文字数の多い順に，部分文字列を抽出するものとして議論する。

さて，*n*-gram 文字列と *m*-gram 文字列の抽出を考える。$n > m$ とすると，条件より，*n*-gram 文字列の抽出は，*m*-gram 文字列に先立って実行される。原文上，*n*-gram 文字列と *m*-gram 文字列が共通部分を持つ場合が問題となるから，それを分類すると，図 **5.2** のように，*m*-gram 文字列が *n*-gram 文字列内に内包される場合と，*m*-gram 文字列と *n*-gram 文字列が互いにその部分を共有する場合に分けられる。

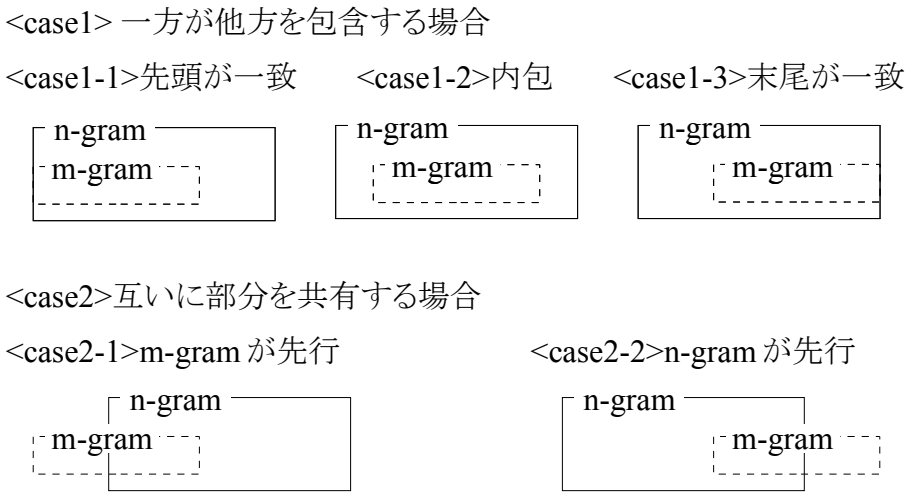


図 **5.2** 抽出文字列の相互関係

34 従来，単語列の場合，結果から計算する方法が使われていた例[73]がある。しかし，その方法では，原文のある一定領域から，互いに部分文字列を共有するような複数の文字列が抽出されているとき，引き過ぎが生じる。抽出の終わった段階では，引き過ぎの有無の判断は下能なため，正確な計算はできない。

(1) 無効化の必要なレコードの範囲

n-gram が先行して抽出されたとき，case1 の *m*-gram は，いずれも抽出対象とならない。したがって，*n*-gram 文字列を抽出するとき，このような関係にある *m*-gram は，後の処理で抽出されないようにする必要がある。そこで，*n*-gram が抽出されたとき，汎用ソートファイル上で，それに包含される *m*-gram を探して，該当レコードが無効とされる条件を付与する方法を考える。

そこで，まず無効化の対象となるレコードについて考えると，case1-1 の場合は，抽出された n-gram のレコード自体が再び抽出の対象にならないようにすればよい。次に，case1-2，case1-3 の場合について考えると，無効化の対象となるレコードは，原文上，着目する *n*-gram の開始文字の位置から数えて *n* 文字先までの各文字を先頭文字とする文字列単語のレコードであることが分かる。

次に，無効化の条件について考えると，case2-2 の場合の *m*-gram は無効化してはならないから，上記の対象レコードのうち，無効化するレコードは，一致文字数がそれぞれ *n*-1，*n*-2，・・・，1 以下のレコードに限られることが分かる。なお，case2-1 にあるような *m*-gram の場合は，上記の無効化処理の対象外となっており，抽出集計の対象となる。

以上の無効化処理の対象範囲について，図 **5.3** に例を示す。図では，原文番地 3 のレコードから 6gram の文字列，「C～H」が抽出対象と判断されたときは，原文番地 4～8 の文字列の H までの部分が無効化されることを示している。

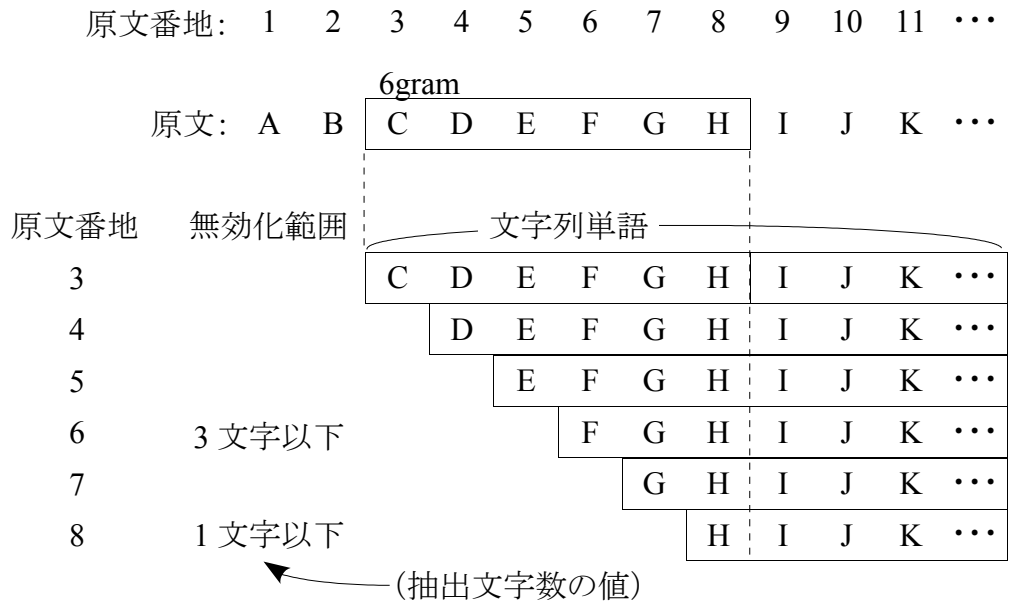


図 **5.3**　文字列採否判定の方法

(2) 無効化すべきレコードの検索

汎用ソートファイルのレコードは，文字列単語を示す原文番地の値 *i* に対して順不同に並んでいる。そのため，あるレコードの原文番地の値 *i* を見て，原文番地の値が，*i*+1，*i*+2，・・・となっているレコードを探すには，シーケンシャルサーチが必要で，検索時間が大きな問題となる。これに対して，元の原文番地ファイルでは，レコードは原文番地の値 *i* の順に並んでいる。すなわち，無効化の要求が発生したレコードに引き続いて，無効化をチェックすべきレコードが順番に並んでいるため，検索は高速に実行できる。そこで，汎用ソートファイルをもう一度，原文番地の値の順に再ソートし，得られたファイル上で無効化処理を行うものとする[35]。

5.3.2 文字列抽出アルゴリズム

(1) アルゴリズム

前節の議論を踏まえ，言語データから，2 回以上の出現回数を持つ固定的な（独立性の高い）表現を文字列として，文字数の多い順に，かつ，重複なしに抽出するアルゴリズムを提案する。

［文字列抽出アルゴリズム］

手順 **1**～手順 **3**：長尾・森の方法と同じ

手順 **4**：「抽出文字数の記入」

汎用ソートファイルの各レコードの示す文字列単語について，先頭から何文字抽出対象となっているか（抽出文字数）を調べ，レコードに記入する（拡張汎用ソートファイルができる）。抽出文字数は，前後のレコードの一致文字数の関係から簡単に決まる。

手順 **5**：「拡張原文番地ファイルの作成」

拡張汎用ソートファイルを原文番号順にソートし直し，拡張原文番地ファイルとする。

手順 **6**：「有効無効判定処理」

拡張原文番地ファイルの各レコードの抽出文字数を順に調べ，各レコードの無効判定を行う。その結果は採否表示の値として記入する。無効判定の方法は，3.1 節で述べたとおりである。

手順 **7**：「再拡張汎用ソートファイルの作成」

上記で得られた拡張原文番地ファイルを再度，汎用ソートファイルのレコード順にソートし，これを再拡張汎用ソートファイルとする。

---

35 汎用ソートファイルの各レコードに，次単語番地(next pointer)のフィールドを設ければ，ランダムアクセスによってたどれる。しかし，通常ファイルサイズは大きく，ディスクアクセス回数が膨大（4 章の実験例では，全レコードに 1 回ずつランダムにアクセスする時間は，1,000 万回×10ms＝30 時間程度と推定される）となる。これに対して連続したレコードの処理（IO バッファのサイズにもよるが）は高速である。そのようにするには，本文で述べたように原文番地順にソートし直す必要があるが，そのためのソート時間は，ソートファイルに次単語番地を探して書き込む処理と同等の時間で実行できる。以上から，ランダムアクセスに比べて，ネクストサーチの方が高速だと期待される。なお，同種の再ソート処理は，連鎖型共起文字列抽出で 2 回，離散型共起で 1 回の合計 3 回必要となるが，いずれも，順不同となったレコード番号が元の連続番号になるようにソートし直すものであり，単純で高速に実行できる（4 章の例では，ソート 1 回当たり数分である）。

手順 **8**:「抽出文字列集計処理」

再拡張汎用ソートファイルの採否表示，抽出文字数，一致文字数の関係を調べて抽出する文字列を決定し，同時に，その出現回数を求める。

このとき，前後のレコードの一致文字数の関係から抽出文字数は求められる（手順4 参照）ため，抽出文字数は参照しなくても集計できる。

(2) 例題検討

以上のアルゴリズムの適用例を図 **5.4** に示す。この例では，*n*-gram 統計で抽出される文字列の種類が 24 種類で，延べ出現回数が 72 回であるのに対して，本節の方法では，5 種類，10 回に絞られる。

5.4 離散型共起表現の抽出法

5.4.1 抽出する共起文字列の条件

2 つ以上の表現が，1 文中の離れた位置に共起するような表現の組（離散型共起表現）と，その出現回数を求める方法を考える。連鎖型共起表現の抽出（5.3 節の方法）では，複数の文にまたがる文字列は抽出の対象外としたため，抽出された連鎖型共起表現は，文内に閉じている。したがって，離散型共起表現を抽出するには，言語データを先頭の文から順にサーチし，連鎖型共起表現の文字列の組が 1 文中に現れる現象を，文字列の組ごとにカウントすればよいが，文境界文字（句点）の扱いと抽出する表現の位置関係が問題となる。

(1) 句点の扱い

通常，日本文は句点で終わるため，句点から句点までを1 文とする。引用文等，1 文内に句点を持つ別の文などを内包する文では，簡単のため，内包される文（対となっている引用記号の区間）は無視する。

(2) 抽出する文字列の相互関係

離散型の文字列共起では，文中で，互いに接続した文字列や部分的にオーバラップする文字列の組は抽出の対象外となる。そこで，5.3 節で抽出された文字列の相互関係について考える。

さて，文字列 α と β が同一の文から抽出された連鎖型文字列とすると，その原文上の位置的関係は，図 **5.5** に示すような 3 つの関係のいずれかとなる。文字列 α と β が分離している(c)の場合は，当然，離散型共起表現の抽出対象になるから，ここでは，(a)，(b)の場合について考える。

［原文データ］ むかし むかしの おかしなおかし。おかしの はなしは おかしなおはなし。

（抽出対象箇所） ―⑤― ―⑤― ―③― ―①― ―②― ―③― ―②― ―④― ―①― ―④―

原文番地ファイル

原文番地	文字列単語［ソートなし］（先頭部分）
1	むかしむかし
2	かしむかしの
3	しむかしのお
4	むかしのおか
5	かしのおかし
6	しのおかしな
7	のおかしなお
8	おかしなおか
9	かしなおかし
10	しなおかし。
11	なおかし。お
12	おかし。おか
13	かし。おかし
14	し。おかしの
15	。おかしのは
16	おかしのはな
17	かしのはなし
18	しのはなしは
19	のはなしはお
20	はなしはおか
21	なしはおかし
22	しはおかしな
23	はおかしなお
24	おかしなおは
25	かしなおはな
26	しなおはなし
27	なおはなし。
28	おはなし。
29	はなし。
30	なし。
31	し。
32	。

［手順**2**］拡張汎用ソートファイルの作成

［手順**3**］一致文字数のカウント

［手順**4**］抽出文字数の記入

拡張汎用ソートファイル

抽出文字数	一致文字数	レコード番号	原文番地	文字列単語［ソートあり］（先頭部分）
5	5	1	8	おかしなおか
5	3	2	24	おかしなおは
3	3	3	16	おかしのはな
3	1	4	12	おかし。おか
1	0	5	28	おはなし。
4	4	6	9	かしなおかし
4	2	7	25	かしなおはな
3	3	8	6	かしのおかし
3	2	9	17	かしのはなし
2	2	10	2	かしむかしの
2	0	11	13	かし。おかし
3	3	12	10	しなおかし。
3	1	13	26	しなおはなし
2	2	14	6	しのおかしな
2	1	15	18	しのはなしは
1	1	16	22	しはおかしな
1	1	17	3	しむかしのお
1	1	18	31	し。
1	0	19	14	し。おかしの
2	2	20	11	なおかし。お
2	1	21	27	なおはなし。
2	2	22	21	なしはおかし
2	0	23	30	なし。
1	0	24	7	のおかしなお
1	0	25	19	のはなしはお
1	1	26	23	はおかしなお
3	3	27	20	はなしはおか
3	0	28	29	はなし。
3	3	29	4	むかしのおか
3	0	30	1	むかしむかし
0	0	31	32	。
0	0	32	15	。おかしのは

［手順**5**］拡張原文番地ファイルの作成

［手順**6**］有効無効判定処理

拡張原文番地ファイル

採否表示	抽出文字数	一致文字数	レコード番号	原文番地	文字列単語
○	3	0	30	1	むかしむかし
×	2	2	10	2	かしむかしの
×	1	1	17	3	しむかしのお
○	3	3	29	4	むかしのおか
○	3	3	8	6	かしのおかし
×	2	2	14	6	しのおかしな
×	1	1	24	7	のおかしなお
○	5	5	1	8	おかしなおか
×	4	4	6	9	かしなおかし
×	3	3	12	10	しなおかし。
×	2	2	20	11	なおかし。お
○	3	1	4	12	おかし。おか
×	2	0	11	13	かし。おかし
×	1	0	19	14	し。おかしの
×	0	0	32	15	。おかしのは
○	3	3	3	16	おかしのはな
○	3	2	9	17	かしのはなし
×	2	1	16	18	しのはなしは
×	1	0	25	19	のはなしはお
○	3	3	27	20	はなしはおか
×	2	2	22	21	なしはおかし
×	1	1	16	22	しはおかしな
○	1	1	26	23	はおかしなお
○	5	3	2	24	おかしなおは
×	4	2	7	25	かしなおはな
×	3	1	13	26	しなおはなし
×	2	1	21	27	なおはなし。
×	1	0	5	28	おはなし。
○	3	0	28	29	はなし。
×	2	0	23	30	なし。
×	1	0	18	31	し。
×	0	0	31	32	。

［手順**7**］再拡張汎用ソートファイルの作成

［手順**8**］抽出文字列集計処理

再拡張汎用ソートファイル

採否表示	抽出文字数	一致文字数	レコード番号	原文番地	文字列単語
○	5	5	1	8	おかしなおか
○	5	3	2	24	おかしなおは
○	3	3	3	16	おかしのはな
○	3	1	4	12	おかし。おか
×	1	0	5	28	おはなし。
×	4	4	6	9	かしなおかし
×	4	2	7	25	かしなおはな
○	3	3	8	5	かしのおかし
○	3	2	9	17	かしのはなし
×	2	2	10	2	かしむかしの
×	2	0	11	13	かし。おかし
×	3	1	12	10	しなおかし。
×	3	2	13	26	しなおはなし
×	2	1	14	6	しのおかしな
×	1	1	15	18	しのはなしは
×	1	1	16	22	しはおかしな
×	1	1	17	3	しむかしのお
×	1	1	18	31	し。
×	2	2	19	14	し。おかしの
×	2	1	20	11	なおかし。お
×	2	2	21	27	なおはなし。
×	2	2	22	21	なしはおかし
×	2	0	23	30	なし。
×	1	0	24	7	のおかしなお
×	1	0	25	19	のはなしはお
×	1	1	26	23	はおかしなお
×	3	3	27	20	はなしはおか
×	3	0	28	29	はなし。
×	3	3	29	4	むかしのおか
×	3	0	30	1	むかしむかし
×	0	0	31	32	。
×	0	0	32	15	。おかしのは

［原文番地］

自レコードが原文の先頭から何番目の文字から始まる単語文字列に対応するかを表示。

［レコード番号］
汎用ソートファイルのレコードの番号。

［一致文字数］
直後のレコードの文字列と比べて先頭から一致する文字数。

［抽出文字数］
自分のレコードで抽出対象となる先頭からの文字数。

［採否表示］
部分文字列として抽出するか否かの判定結果。

抽出された文字列

方法	提案方法		従来方法	
gram 数	文字列	出現	文字列	出現
5gram	① おかしなお	2	おかしなお	2
4gram	― ― ― ―	―	んかしな	2
	― ― ― ―	―	かしなお	2
3gram	② おかし	2	おかし	4
	― ― ―	―	かしな	2
	③ かしの	2	かしの	2
	― ― ―	―	しなお	2
	④ はなし	2	はなし	2
	⑤ むかし	2	むかし	2
2gram	― ―	―	おか	4
	― ―	―	かし	6
	― ―	―	しな	2
	― ―	―	しの	2
	― ―	―	なお	2
	― ―	―	なし	2
	― ―	―	はな	2
	― ―	―	むか	2
1gram	―	―	お	5
	―	―	か	6
	―	―	し	8
	―	―	な	4
	―	―	の	2
	―	―	は	3
	―	―	む	2
― ― ―	合計	10	合計	72

2 回以上出現した文字列で、包含されていない文字列

図 **5.4**　連鎖型共起表現抽出アルゴリズム実施例

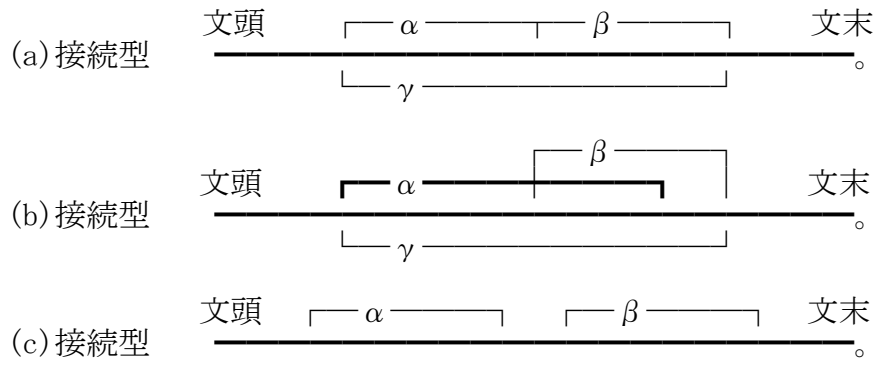


図 **5.5** オーバラップした文字列の扱い

(a) 文字列？と？が接続している場合

言語データ中，このような文字列を含む場所は，最大 1 カ所である。なぜなら，そのような文字列を含む場所が 2 カ所以上ある場合は，文字列 αβ がより文字数の多い文字列 γ として集計され，それらの文中の部分文字列 α および β はカウントされないからである。したがって，文字列 α と β が文中に共起する回数が 2 回以上ある場合は，最大 1 文を除く他の該当する文は(c)のタイプ（分離型）の共起となっている。この場合，(a)のタイプの共起は，通常，分離型で共起する文字列がたまたま接続したものとみなせるから，離散型共起表現の抽出対象となる。

(b) 文字列？と？がオーバラップしている場合

文字列 α と β を包含する文字列を γ とする。前項と同様，このような文字列 γ が，言語データ内に 2 カ所以上出現した場合は，γ 自身が連鎖型共起表現の抽出の対象となり，その部分に含まれた文字列 α と β は，抽出されない。したがって，原文中，(b)のような関係にある文字列 α と β が抽出された文は，高々 1 文に限られ，α と β が共起する残りの文は，いずれも(c)のタイプの共起である。しかし，この場合は，(b)のタイプの α と β は文中の共起とは言えないから，抽出集計の対象とならない。

以上から，文内の離散型共起表現の抽出においては，(b)のタイプの共起のみを抽出対象外とすればよい。

(3) 表現要素の出現順序の扱い

離散型共起表現では，それを構成する表現要素（ここでは，連鎖型共起表現として抽出された部分文字列）の出現順序は意味を持つため，出現順序を区別して抽出し集計する。

5.4.2 文字列抽出アルゴリズム

(1) アルゴリズム（図 5.6 参照）

［前準備］

再拡張汎用ソートファイル上の連鎖型共起表現として抽出された文字列に文字列番号を付与する。

手順 **9**:「再拡張汎用ソートファイルの再ソート」

再拡張汎用ソートファイルを原文番地の値の順にソートし，拡張原文番地ファイルのレコード順に戻す。

手順 **10**:「文番号の付与」

得られたファイルの各レコードに文番号を記入する。

手順 **11**:「ファイルの圧縮」

上記ファイルを以下の手順で圧縮し，「離散型共起圧縮ファイル」を作成する（次の手順に備えて，不要な作業領域を開放する）。

① 文番号，文字列番号，抽出文字数，原文番地の 4 つの欄以外は，削除する。

② 文字列番号の欄の値がないレコードを削除する。

手順 **12**:「離散型共起表現の抽出とカウント」

一般に，k:種類（$k \geqq 2$）の文字列からなる離散型共起表現を抽出するものとすると，同一の文内にある文字列番号の k 個の組み合わせのすべてを（文中の出現順序の順にセットにする）ファイルに書き出し，それをソートして，同一の組の数をカウントする。

以上で，離散型共起表現の集計表が求められる。これらの表現を含む文を出力するには，手順 12 で作成する各表現の組に文番号を追記しておけばよい。

(2) 例題検討

以上の手順を，図 **5.4** の例に適用し，要素数 2 の離散型共起表現を求めた。その結果を図 **5.6** に示す。この例では，5.3 節で抽出された 5 種の連鎖型共起文字列 25 組中，1 文内に離れて 2 回以上，共起する文字列の組が 6 組で，それらの延べ出現回数は 12 回である。

［原文データ］ むかし むかしの おかしなおかし。おかしの はなしは おかしなおはなし。

（抽出対象箇所）

再拡張汎用ソートファイル（文字列番号付き）

文字列番号	採否表示	抽出文字数	一致文字数	レコード番号	原文番地	文字列単語
①	○	5	5	1	8	おかしなおか
①	○	5	3	2	24	おかしなおは
②	○	3	3	3	16	おかしのはな
②	○	3	1	4	12	おかし。おか
	×	1	0	5	28	おはなし。
	×	4	4	6	9	かしなおかし
	×	4	2	7	25	かしなおはな
③	○	3	3	8	5	かしのおかし
③	○	3	2	9	17	かしのはなし
	×	2	2	10	2	かしむかしの
	×	2	0	11	13	かし。おかし
	×	3	3	12	10	しなおかし。
	×	3	1	13	26	しなおはなし
	×	2	2	14	6	しのおかしな
	×	2	1	15	18	しのはなしは
	×	1	1	16	22	しはおかしな
	×	1	1	17	3	しむかしのお
	×	1	1	18	31	し。
	×	1	0	19	14	し。おかしの
	×	2	2	20	11	なおかし。お
	×	2	1	21	27	なおはなし。
	×	2	2	22	21	なしはおかし
	×	2	0	23	30	なし。
	×	1	1	24	7	むかしのおか
	×	1	0	25	19	むかしむかし
	×	1	1	26	23	。
④	○	3	3	27	20	。おかしのは
④	○	3	0	28	29	
⑤	○	3	3	29	4	
⑤	○	3	0	30	1	
	×	0	0	31	32	
	×	0	0	32	15	

図 **5.4** の［手順 7］から続く

［前処理］再拡張汎用ソートファイルに加工

［手順 **9**］再拡張汎用ソートファイルの再ソート

［手順 **10**］文番号の付与

拡張原文番地ファイル（文番号付き）

文番号	文字列番号	採否表示	抽出文字数	一致文字数	レコード番号	原文番地	文字列単語
1	⑤	○	3	0	30	1	むかしむかし
1		×	2	2	10	2	かしむかしの
1		×	1	1	17	3	しむかしのお
1	⑤	○	3	3	29	4	むかしのおか
1	③	○	3	3	8	5	かしのおかし
1		×	2	2	14	6	しのおかしな
1		×	1	1	24	7	のおかしなお
1	①	○	5	5	1	8	おかしなおか
1		×	4	4	6	9	かしなおかし
1		×	3	3	12	10	しなおかし。
1		×	2	2	20	11	なおかし。お
1	②	○	3	1	4	12	おかし。おか
1		×	2	0	11	13	し。おかしの
1		×	1	0	19	14	。おかしのは
1		×	0	0	32	15	おかしのはな
2	②	○	3	3	3	16	しのはなしは
2	③	○	3	2	9	17	のはなしはお
2		×	2	1	15	18	はなしはおか
2		×	1	0	25	19	なしはおかし
2	④	○	3	3	27	20	はおかしなお
2		×	2	2	22	21	おかしなおは
2		×	1	1	16	22	かしなおはな
2		×	1	1	26	23	しなおはなし。
2	①	○	5	3	2	24	なおはなし。
2		×	4	2	7	25	おはなし。
2		×	3	1	13	26	はなし。
2		×	2	1	21	27	なし。
2		×	1	0	5	28	し。
2	④	○	3	0	28	29	。
2		×	2	0	23	30	
2		×	1	1	18	31	
2		×	0	0	31	32	

［手順 **11**］ファイルの圧縮

離散型共起圧縮ファイル

文番号	文字列番号	抽出文字数	原文番地
1	⑤	3	1
1	⑤	3	4
1	③	3	5
1	①	5	8
1	②	3	12
2	②	3	16
2	③	3	17
2	④	3	20
2	①	5	24
2	④	3	29

［手順 **12**］離散型共起表現の抽出とカウント

離散型表現吐き出しファイル

(⑤, ⑤)
(⑤, ③)
(⑤, ①)
(⑤, ②)
(⑤, ③)
(⑤, ①)
(⑤, ②)
(③, ①)
(③, ②)
(①, ②)
(②, ③)
(②, ④)
(②, ①)
(②, ④)
(③, ④)
(③, ①)
(③, ④)
(④, ①)
(④, ④)
(①, ④)

ソート

(①, ②)
(①, ④)
(②, ①)
(②, ③)
(②, ④)
(②, ④)
(③, ①)
(③, ①)
(③, ②)
(③, ④)
(③, ④)
(④, ①)
(④, ④)
(⑤, ①)
(⑤, ①)
(⑤, ②)
(⑤, ②)
(⑤, ③)
(⑤, ③)
(⑤, ⑤)

離散型共起表現集計表［最終結果］

前方の文字列	文内後方の文字列と出現回数
② おかし	④ はなし 2回
③ かしの	① おかしなお 2回　④はなし 2回
⑤ むかし	① おかしなお 2回　②おかし 2回　③かしの 2回

図 **5.6**　離散型共起表現抽出アルゴリズム実施例（図 **5.4** の手順 **7** から続く）

5.5 共起表現の抽出実験

本章で提案した方法の効果を検証するため，日本語データへの適用例として，日経新聞記事3カ月分(892万字)を対象に，連鎖型共起文字列および離散型共起文字列の抽出実験を行った。ただし，読点を除く記号類を含む文字列は抽出の対象としないこととした。

5.5.1 連鎖型共起表現の抽出

抽出する文字数もしくは文字列の出現回数を制限した場合に，抽出される文字列の種類数と延べ出現回数を従来の方法と比較して，表 **5.1**，表 **5.2** に示す。文字列の長さから見た，抽出される文字列の種類数，出現回数および文字列の例を表 **5.3** に示す。また，出現頻度の高い文字列の例を表 **5.4** に示す。

表 **5.1** 抽出された文字列の種類と延べ度数(その **1***)

結果比較 / 抽出対象	本章 の方法		長尾・森 の方法		減少の 結果	
	a) 文字列の種類数	b) 延べ出現回数	c) 文字列の種類数	d) 延べ出現回数	a/c	b/d
2 文字以上	970,203	2,613,704	4,374,141	31,178,897	22.2 %	8.38%
5 文字以上	591,901	1,476,922	2,960,487	10,808,458	20.0 %	13.7 %
10 文字以上	52,214	114,270	673,601	1,550,817	7.75%	7.37%
20 文字以上	1,792	3,692	177,298	359,810	1.01%	1.03%

* 文字列の長さから見た集計

表 **5.2** 抽出された文字列の種類と延べ度数(その **2***)

結果比較 / 抽出対象	本章 の方法		長尾・森 の方法		減少の 結果	
	a) 文字列の種類数	b) 延べ出現回数	c) 文字列の種類数	d) 延べ出現回数	a/c	b/d
2 回以上	970,203	2,613,704	4,377,087	39,588,291	22.2 %	6.60%
5 回以上	67,321	551,441	882,217	31,288,701	7.63%	1.76%
10 回以上	12,351	217,934	372,291	28,050,199	3.32%	0.78%
20 回以上	2,288	92,804	169,375	25,871,964	1.35%	0.36%
50 回以上	285	37,850	62,991	22,209,875	0.45%	0.17%
100 回以上	76	24,167	30,316	19,961,961	0.25%	0.12%
200 回以上	20	16,771	14,363	17,759,432	0.14%	0.07%

* 出現頻度から見た集計

表 **5.3** 抽出された文字列の例（出現回数の多い順に掲載：**( )**内は出現回数）

文字数	本章の方法	長尾・森の方法
2文字	また(325), 写真(315), 東京(281), 価格(276), 連帯(198), など(198), ただ(198), この(180), 時事(178), さん(162), 一、(143), EC(132), 共同(129), 当時(125), 特に(117), だが(117), 二五(114), とか(110), 私は(104), 政府(99) ［合計40,843種類 延べ165,177回］	てい(55025), して(47026), いる(45236), する(39828), した(36679), って(31790), った(29370), こと(27816), から(27428), ない(26396), ある(24223), など(22126), が、(20855), は、(20631), とい(19388), ると(7654) ［合計1,413,654 種類延べ7,954,309回］
5文字	としている(436), 欧州共同体(277), このため、(158), 市場占有率(141), とみている(141), モーターズ(133), と強調した(130), これに対し(126), この結果、(112), 国民総生産(110), このほか、(107), ところが、(105), を発表した(92), それだけに(89), その結果、(89) ［合計190,925種類 延べ499,653回］	なっている(3710), ているが、(2827), によると、(2753), については(2721), されている(2334), ることにな(2286), になってい(2079), としている(1997), 五十七年度(1849), ボーランド(1818), ていること(1776), しており、(1697), することに(1581), となってい(1544), 明らかにし(1517) ［合計748,172種類 延べ3,793,077回］
10文字	することになっている(44), 第二次臨時行政調査会(35), することになりそうだ(19), 82ジャパンショップ(17), しているのではないか(16), ワシントン一九日共同(14), サウジアラビア通過庁(14), として注目されている(14), ニューヨーク二十五日(13), を余儀なくされている(13) ［合計21,155種 延べ47,336回］	したところによると、(273), らかにしたところによ(223), 明らかにしたところに(223), かにしたところによる(222), にしたところによると(222), 第二次臨時行政調査会(208), ることを明らかにした(191), 日明らかにしたところ(173), たことを明らかにした(120), ホテルニュージャパン(112) ［合計132,865種類 延べ345,232回］
20文字	参院全国区制改革のための公職選挙法改正案(5), 五十九年一月から実施されるグリーンカード(5), 鈴木首相は一六日午後の参院予算委員会で、(4), 本社東京、社長林健彦氏、資本金百七十億円(4), 十月一日会社訪問、十一月一日入社試験解禁(4) ［合計823種 延べ677回］	の最も多かった中心相場は前年中心値に比べ(18), 引の最も多かった中心相場は前年中心値に比(18), 取引の最も多かった中心相場は前年中心値に(18), の国立オリンピック記念青少年総合センター(17), 木の国立オリンピック記念青少年総合センタ(17) ［合計14,625種類 延べ30,664回］

表 **5.4** 出現頻度の高い文字列の例

頻度の高い文字列 200回以上	という(586), と述べた(512), としている(436), また(325), である(324), 写真(315), しかし、(302), と語った(283), 東京(281), 価格(278), 欧州共同体(277), しかし(274), ポイント(269), ひとこと(264), 発売期間(259), また、(236), これは(220), このため(204), ただ、(201)

これらの表から，以下のことが分かる。

本章の方法では，期待されたとおり，従来の方法に比べて，多くの断片的な文字列の抽出が抑制され，抽出される文字列の種類，出現回数共に大幅に減少する。例えば，2 文字以上，2 回以上の文字列では，抽出される種類が約 5 分の 1，延べ出現回数も約 12 分の 1 に抑制される。この効果は，文字数の大きい文字列ほど大きく，20 文字以上の場合では，抽出される文字列の種類，延べ出現回数共に，約 100 分の 1 になる。

5.5.2 離散型共起文字列の抽出

(1) 抽出文字列の性質

簡単のため，単独ではそれぞれ 10 回以上出現した 2 種類の文字列が 1 文内に離れて共起する場合[36]について，抽出された文字列の組の数を表 **5.5** に示す。また，出現頻度の多い文字列の組と，2 回以上出現した文字列の組の中で合計文字数の多い文字列の組を，それぞれ，表 **5.6**，表 **5.7** に示す。

表 **5.5** 抽出された文字列の組の種類と延べ度数

結果 / 抽出対象	文字列の組の種類数	延べ出現回数
2 回以上	6,544	21,829
5 回以上	941	9,057
10 回以上	237	4,556
20 回以上	61	2,291

（2 種類の文字列の文内共起の場合）

表 **5.6** 出現頻度の高い文字列の組の例

順位	前方文字列	後方文字列	回数
1 位	価格	発売時期	257
2 位	ゼネラル	モーターズ	117
3 位	サミット	先進国首脳会議	86
4 位	EC	欧州共同体	80
4 位	イラン	ジャパン石油化学	80

表 **5.7** 合計文字数の大きい文字列の例

順位	前方文字列	後方文字列	合計字数	回数
1 位	部会長、梅本純正武田薬品工業副社長	について協議した	25	2
2 位	本社夕張市、保全管理人山根喬氏	問題などについて	23	2
3 位	北炭夕張炭坑	本社夕張市、保全管理人山根喬氏	21	9
4 位	永田町のホテルニュージャパン	横井英樹社長	21	11
5 位	マルクは一ドル	フラン、英ポンドは一ポンド	20	5

36 2 種類の表現の組の集計では，足切りをしない（2 度数以上を対象とする）場合，約 18 万種類（延べ 40 万度数）の離散型表現が抽出された。ここでは，結果を見やすくするため，抽出される種類が約 1 万件以下になるように，単独出現回数 10 で，入力の足切りをした場合を示す。

表 **5.6**, 表 **5.7** から, 出現頻度の高い離散型共起の多くは, 名詞同士の共起であることが分かる。特に, 話題として新聞記事に取り上げられた固有名詞や日時等の数量との共起が数多く取り出されている[37]。

このような名詞の共起情報は, 例えば, 機械翻訳用の辞書作成などに応用できる。また, テンプレート翻訳などでは, 名詞同士の共起よりもむしろ, 文型パターンを作り易い助詞や助動詞を含む表現要素の共起を収集したい場合がある。表 **5.5** を見ると, 抽出された表現の組は, すでにかなり絞り込まれているため(全体で, 6,544 件), 全体を人手によってチェックし, 助詞, 助動詞を含む表現の組など, 目的に応じた表現の組を選択して取り出すことはさほど困難ではない。

しかし, さらに大量の言語データの場合, 出力されるデータ量が増大し, 人手による選択が困難となることが考えられる。そのような場合, 得られた結果から目的にあわないような表現を選択的に削除する方法もあるが, 連鎖型共起, 離散型共起の抽出処理の過程に介入して, 抽出対象文字列に制限を加えることもできる。なるべく早い段階で, 抽出対象とする文字列の字種構成に制約を加えたり, 抽出された文字列をチェックして, 不要なものを削除したりすれば, その後の計算量は減少し, 出力結果のチェック作業も減少する[38]。

ここでは一例として, ひらがな文字を含まない文字列と記号英数字を含む文字列は抽出しないという条件で得られた離散型共起の結果の一部を表 **5.8**, 表 **5.9** に示す。この場合, 新聞記事の文型に相当するような, 離散型共起表現が抽出されることが分かる。

表 **5.8** 抽出された離散型共起表現の例(**( )**内は出現回数)

～としながらも～と述べた(9), ～の質問に答え～と述べた(9), その内容は～というもの(6),
われわれは～と語った(6), さらに首相は～と述べた(5), その内容は～など(5),
～とし、～と述べた(5), ～についても～としている(4), いかにも～らしい(4), つまり～である(4),
～にしろ～にしろ(4), このほか～などと語った(4), これに対し～と答えた(4),
この中には～も含まれている(3), これに対し～と反論した(3), これに対して～と答えた(3),
～するつもりだ～と語った(3), これからは～という(3), この骨子は～というもの(3),
～にせよ、～にせよ(3), その内容は～などとなっている(3), ～などから～とみている(3),
～にするか～にするか(3), ～と述べるとともに～と語った(3), ～なり～なり(3),

(ひらがなを含み, 記号英数字を含まない要素を抽出した結果)

37 表 **5.6** では「ゼネラル」+「モータース」, 「サミット」+「先進国首脳会議」などのペアか頻出しているが, これは, 本文中には「ゼネラル・モータース」, 「サミット(先進国首脳会議)」などとして出現していたためで, 読点を除く記号類は連鎖型共起の集計の対象としなかったためである。

38 例えば, 連鎖型で抽出した文字列の 10%が有効な表現だったとすると, 離散型の場合に抽出される文字列の組の有効なものは, 0.1 の n 乗 (n は要素とする表現の数) 以下に減少すると考えられる。したがって, 連鎖型に比べて離散型ではさらに, 抽出したい表現をいかに絞り込むかが重要な問題となる。実験によれば, 抽出したい表現を字種によって制約する効果が大きいが, この点は, さらに今後の検討が必要である。

表 **5.9** 合計文字数の多い離散型共起表現の例

前方文字列	後方文字列	合計字数	回数
することになろう	との見通しを明らかにした	20	2
その結果	との意見が大勢を占めた	16	2
しているうえ、	していることから	16	2
についても	することになりそうだ	15	2
と語り、	する方針を明らかにした	15	2
であれ、	することになりかねない	15	2
その理由として	などをあげている	15	2
しかし、こうした	が出てきている	15	2

（ひらがなを含み，記号英数字を含まない要素を抽出した結果）

(2) 言語データ量と処理サイズについて

離散型共起の場合は，連鎖型共起で抽出した表現の組を扱うため，表現の組を書き出すためのファイルの容量が問題となると予想される。このファイルの必要量は，離散型共起として生起した表現の数（頻度 1 以上の延べ度数）で決まる。

実験例によれば，連鎖型共起で抽出した表現 97 万種類（延べ度数 260 万回）を足切りをせず（ただし度数 1 のものは除く），そのまま使用して要素数 2 の離散型共起を計算すると，度数 2 以上の離散型表現として，18 万種類（延べ度数 40 万回）の表現の組が得られた。このとき，手順 12 でファイルに書き出された表現の組（ただし文字列番号のペア）は，2,000 万組で，それに要したファイル量は 400MB（20 バイト/文字列ペア）であった。

これに対して，連鎖型共起として抽出された文字列のうち，度数 10 以上のもの 1.2 万種類（延べ度数約 22 万回）を取り上げ，それらを要素とする離散型共起表現を求めた場合は，2 度数以上の離散型共起表現として，6,500 種類（延べ度数約 2 万回）の表現が得られた。この計算の過程でファイルに書き出された表現の組は，約 58 万組で，使用したファイル量は約 12MB であり，足切りをしない場合に比べて，1/30 以下に減少した。

ここで，言語データ量と処理サイズの関係を考える。連鎖型共起で抽出される表現の延べ度数は，言語データ量にほぼ比例し，離散型共起で抽出される表現の延べ度数は，連鎖型共起で得られた表現の延べ度数の 2 乗にほぼ比例すると考えられろから，離散型共起集計用のファイル使用量は，言語データ量の 2 乗にほぼ比例すると推定される。しかし，言語データ量が増加したときは，それに比例して連鎖型共起表現の足切り値を上げても抽出精度は低下せず，重要な（頻度の高い）表現は漏れなく収集できると期待される[39]。そこで，表 **5.2** を見ると，足切り値にほぼ反比

39 共起表現の抽出では，出現頻度の高い表現をいかにもれなく拾い出すかが問題である。出現する表現の分布に大きな偏りのない標本であれば，標本量を増加させたとき，それにつれて出現頻度の高い表現の出現回数も増加するから，適当な値で足切りをしてもそれらを漏らす心配は少ない。なお，表現に大きな偏りのある標本の場合は，ジャンルごとに分けて，共起表現を収集する方が適切と言える。

例して，抽出される連鎖型共起表現の延べ度数は減少していることが分かる。

これらの点から，原文データ量が増加したときは，足切り値をそれに比例して上げることにより，抽出精度を低下させないで離散型共起の計算ができ，そのとき，計算に必要とされるファイル量の増加は，言語データの増加に比例するオーダに抑えられると期待できる。

5.5.3 今後の改良と応用について

(1) 目的に合わせた抽出文字列種別の指定

離散型の共起表現抽出の場合，計算可能な言語データ量を増大させるためには，特に，それに使用する連鎖型文字列の種類を少しでも減少させることが望まれる。これに対して，実験例で抽出された文字列には，まだ，様々な種類の文字列が混ざっている。日本語の場合，例えば，

① 数字，カタカナ語，英字略語からは，多くの断片的な文字列が抽出されやすい。
② 文型を決めるようなキーワードはかな文字を 1 文字以上含む場合が多い。
③ 漢字を含まない複合名詞は少ない。

などの性質に着目し，手順 3 以降で，抽出する文字列を構成する字種を指定すれば，不要な文字列の抽出はさらに抑制できる。

(2) 抑制された文字列カウントの一部復活

本章では，連鎖型共起の計算において，一度抽出した文字列内の部分文字列の抽出はダブルカウントになると考え，条件 3（最長一致のもののみ抽出）を前提とした。このため，抽出される文字列は，独立性があり，連鎖共起とみなせる文字列に絞られている。しかし，より細かい要素からなる離散型共起をも収集しようとする場合は，一度抽出した文字列の中の要素からも，要素的な表現を抽出すればよい。

断片的な文字列の抽出を抑制しながら，これらの要素的表現を抽出するには，図 **5.1** で，文字列 α の中に含まれる部分文字列の β や γ も，その文字列が原文中の他の部分に生起して抽出対象となったときはカウントに加えると良い[80]。具体的には，5.3 節のアルゴリズムの手順 8 で，*n*-gram の文字列を抽出する際，そのレコードの上方向に連続するレコードで，抽出文字数の値が *n*+1 以上のものも *n*-gram の抽出対象に加えればよい。その際，新たに抽出対象となったレコード（重複抽出の対象レコード）をコピーして追加すれば，離散型共起の計算処理の手直しは不要となる。

(3) 離散型共起表現抽出における 1 文字要素の扱い

本節の実験では，計算量を減少させるため，抽出対象文字列の文字数は 2 文字以上であるとした。しかし，離散型共起表現の抽出において，日本文の文型を抽出したいような場合，「～が～を～に～」などのように，文中から 1 文字キーワードの組を探したい場合がある。このような場合は，

後に述べるように，形態素解析結果に対して，本章の方法を適用すればよいと考えられるが，(1)で述べた方法などにより，抽出対象を絞り込むことによって計算量を減らし，抽出を可能とすることも考えられる。

(4) 形態素列，単語列等への適用

日本語の文型を抽出するには，言語データを形態素解析して得られた単語の文法的属性や意味属性を表す記号列に対して，本章の方法を適用することが期待される。文法的，意味的に見てどのような種類の文型情報が得られるか，また，単語共起情報を得る場合，文字連鎖に適用する方法と，単語列に適用する方法のどちらがよいかなど，今後の課題である。

5.6 結言

言語コーパスなどの膨大な言語データの中から，使用頻度の高い表現および表現の組を自動的に発見し集計する方法を提案した。具体的には，まず，任意の *n*-gram の計算法として提案された長尾らのアルゴリズムを独立性の高い表現を抽出する観点から改良し，言語データの中に2回以上出現した文字列（連鎖型共起表現）を，「一度，抽出した文字列の部分文字列は，その後，抽出対象としない」という条件下で，漏れなく自動的に抽出し集計する方法を提案した。次に，この方法で抽出された文字列を組み合わせて，文中の離れた位置に共起する文字列の組（離散型共起表現）を抽出し，その頻度を求める方法を示した。

3カ月分の新聞記事データ（892万字）に適用した例によれば，連鎖型共起表現抽出の場合，従来の方法では，2文字以上，2度数以上の文字列が，440万種類，延べ3,120万回の文字列が抽出されたのに対して，本章の方法では，97万種類，延べ260万件に減少した。抽出された文字列を比較した結果，*n*-gram の方法で得られた文字列が，膨大な量の断片的な文字列（文法的，意味的に意味のない文字列）を含むのに対して，本章の方法では，それらの断片的な文字列が大幅に削除されることが確認された。

この効果により，離散型の共起表現の網羅的な自動抽出が可能となった。提案した離散型共起表現抽出方式の適用例では，連鎖型共起の集計で得られた文字列のうち，10回以上出現した文字列（12,350種類）の任意の2種類が，1文中に2回以上共起した表現の組は，6,500種類（延べ出現回数21,800回）であることなど，離散型の共起表現が容易に求められることが分かった。

以上のとおり，本章の方法では，連鎖型共起表現抽出での断片的文字列の抽出が抑制される結果，離散型共起表現を容易に計算することが可能となり，文型パターンなど，文構造に関する基礎データを，ほぼ自動的に収集することが可能となった[40]。

40 n-gram の方法はデータ分析の方法として広く利用され，様々な改良が進められている。人文系でも特徴分析の手がかりを得るために利用されている[81]。離散型の共起表現の網羅的な抽出はあまり行なわれていないが，Sander らが suffix array を用いることにより本章の方法より計算量を削減する方法を提案している[82]。

第6章　結論

本論文では，日英翻訳の精度を向上させることを目的として，ルール型翻訳，用例型翻訳の精度向上に有効な方法を検討した。本論文により得られた成果をまとめると次のとおりである。

第 2 章では，長文解析精度の低下要因であった述語間の係り受け関係の曖昧さを解決するため，日本語の意味的な階層的表現構造に着目した，従属節間の係り受け解析方式を提案した。具体的には，日本語表出過程に着目した南の 3 段階の階層的な従属節分類を見直し，意味と形式に着目して，基本分類 13 種，細分類 4 種に詳細化し，それらの係り述節，受け述節としての関係を分類整理することにより，述語間の係り受け関係を決定する方法を提案した。新聞記事 972 文（述語数合計 2,327 件，そのうち係り受け曖昧述語 661 件）を対象にした評価によれば，従来の方法では係り先の曖昧な述語が 356 件残ったのに対して，提案した方法では 54 件に減少した。文単位に見れば，述語間の関係が一意に決定できる文の割合は 73.2%から 94.4%に向上した。並列構造解析については黒橋らが有効な方法を提案していることをあわせると，係り受け解析の 2 大問題（並列構造の解析，述語間の関係解析）が解決に向けて大きく前進した。

第 3 章では，機械翻訳の品質を向上させるための 1 つの方法として，(1)精密な単語意味属性を使用して書き替え規則を記述すること，(2)書き替え規則適用条件の判定可能な情報が得られる構文解析結果に規則を適用すること，によって副作用の少ない原文自動書き替え型の翻訳方式を実現した。新聞記事 102 文に対する翻訳実験によれば，書き替え規則の適用された箇所は 102 文中，44 文，延べ 52 箇所で，訳文品質向上効果のあった文は 33 文であった。また，適用された文の構文意味解析の多義の数が平均 5.4/文から 1.3/文まで減少した。提案方式は，翻訳品質向上，多義解消の双方において大きな効果がある。インプリメントの観点からみても，提案本方式は「翻訳困難な表現の翻訳に，既存の翻訳機能がそのまま利用できる」点で，大きな利点があり，今後の訳文品質向上策として有望であると判断できる。

第 4 章では，日英機械翻訳において，用言（動詞，形容詞）の意味を訳し分けるのに必要な結合価パターン対の数とそれを収集する手段について検討した。具体的には，単語当たりの語義が多いためパターン対作成が最も困難な和語動詞の場合を取り上げ，(1)和英辞書から収集する方法，(2)日本語動詞の語義対応の用例を使用する方法，(3)それらを参考に，人の知識に基づいて用例を作成して使用する方法の 3 種のパターン対の収集方法を比較した。その結果，主要な約 1,000 の和語動詞を意味によって訳し分けるには 7,500 件の結合価パターンが必要であることが分かった。これに対して，従来の和英辞書から収集できるパターン対の数は約 1/4，和英辞書と日本語辞書の語義分類知識を使用する場合は約 1/2 であること，必要なパターン対を網羅的に収集するには，作業工数の面でも，和英辞書と日本語辞書の語義を参考に人の知識を内省して用例を作成する方法が適していることなどが分かった。また，この結果から推定すると，漢語動詞，形容詞系の述語，用言性慣用表現などを含むパターン対全体では約 25,000 パターンが必要なこ

と，それらのパターンも辞書等を参考に人の知識を内省する方法で抽出された用例から比較的容易に収集できる見込みであることが分かった。このようにして作成された結合価パターン対辞書に基づいて，日本語語彙大系の構文体系は作成されている。

第 5 章では，膨大な言語データの中から，使用頻度の高い表現および表現の組を自動的に発見し集計する方法を提案した。具体的には，任意の *n*-gram の計算法として提案された長尾・森のアルゴリズムを改良し，言語データの中に 2 回以上出現した文字列（連鎖型共起表現）を，「一度，抽出した文字列の部分文字列は，その後，抽出対象としない」という条件下で漏れなく抽出する方法を提案した。次に，この方法で抽出された文字列を組み合わせて，文中の離れた位置に共起する文字列の組（離散型共起表現）を抽出する方法を示した。新聞記事データ 3 カ月分（892 万字）に適用した実験によれば，連鎖型共起表現は，従来方法では，2 文字以上，2 度数以上の文字列が，440 万種類，延べ 3,120 万回の文字列が抽出されたのに対して，提案方法では，97 万種類，延べ 260 万件に減少した。このとき，提案方法では，断片的な文字列（文法的，意味的に意味のない文字列）が従来方法に比べて大幅に減少することが確認された。提案方法をさらに進めて，離散型の共起表現の網羅的な自動抽出方法を提案した。提案方法の適用例では，連鎖型共起の集計で得られた文字列のうち，10 回以上出現した文字列（12,350 種類）の任意の 2 種類が，1 文中に 2 回以上共起した表現の組は，6,500 種類（延べ出現回数 21,800 回）であることなど離散型の共起表現が容易に求められることが分かった。この結果，提案方法を用いることにより，文型パターンなど，文構造に関する基礎データを，ほぼ自動的に収集することができるため，テンプレート翻訳のルール作成の大幅な効率化が可能になった。

本論文で提案した方式はそれぞれ日英翻訳の精度向上に有効であるが，これらを有機的に結合することによりさらに精度を向上させることが可能になると考えられる。今後の課題としては，この有機的結合方法の検討が必要である。一方，実用性の観点からは，方式提案だけでなく，その方式が必要とする言語知識やデータベースの実現性を併せて検討することが重要である。例えば，用例翻訳においては数多くの方式が提案されているが，必要とされる対訳コーパスの実現性に疑問があるものが少なくない。新聞記事を利用すれば緩く対応付けられた対訳コーパスが実現可能であると考えられるので，それを利用した用例翻訳は実用に耐える可能性がある。用例利用型翻訳方式の提案として付録 A で述べる。

また，本論文では，言語的な性質が大きく異なる日本語から英語への翻訳に的を絞って検討したが，他の言語間の翻訳に適用できる技術も少なくないと考えられる。しかし，言語的な性質が大きく異なる言語間の翻訳技術の確立には日英翻訳と同様の言語リソース量や人的，時間的コストを必要とするため，日英翻訳と同様の手順を積み上げることによって多言語翻訳を実現するのは現実的なアプローチではないと考えられる。日韓や英仏などの言語的な性質が類似している言語間の翻訳は，比較的容易に高精度の翻訳が得られていることを勘案すれば，言語的な性質を類型化して，類型ごとに代表言語を 1 つ選択し，相互に言語的性質が大きく異なる代表言語間の翻訳を実現すること，言語的性質が類似する類型内の言語間翻訳を実現することという 2 つに分けて機械翻訳を検討することが効果的であると考えられる。今後は，類型内の言語間の翻訳の実現についても検討を進めたいと考えている。

謝辞

本論文をまとめるにあたり，懇切なるご指導，ご教示を賜った東京工業大学大学院理工学研究科の田中穂積 教授，徳永健伸 助教授 に心から感謝します。また，本論文に対して有益なご意見，ご助言を賜った同大学 古井貞熙 教授，佐伯元司 教授，奥村学 助教授に深く感謝します。

本研究は，筆者がNTT 電気通信研究所に在籍中に，多数の方々のご協力を得て行なわれたものです。NTT コミュニケーション科学研究所の河岡司 元所長（現在，同志社大学教授）と松田晃一 元所長（現在，NTT アドバンステクノロジ（株）常務）には，本研究の機会を与えていただき，研究途上においてもご指導，ご援助をいただきました。NTT コミュニケーション科学研究所の八巻俊文 元研究部長と大山芳史 元研究部長（ともに現在，NTT アドバンステクノロジ（株））には，本研究の推進にあたり多大のご指導，ご援助をいただきました。NTT コミュニケーション科学研究所の池原悟 元研究グループリーダ（現在，鳥取大学教授）には，本研究の初期から直接ご指導賜りました。心から御礼申し上げます。日英翻訳研究グループのメンバであった宮崎正弘 氏（現在，新潟大学教授），横尾昭男 氏（現在，NTT アドバンステクノロジ（株）），林良彦 氏（現在，大阪大学教授），奥雅博 氏，中岩浩巳 氏ほか，小倉健太郎 氏，菊井玄一郎 氏（現在，（株）国際電気通信基礎技術研究所 室長），松尾義博 氏，Francis Bond 氏，畑山満美子 氏（現在，東日本電信電話（株））には，ルール型翻訳の研究やその技術の応用に関してご協力とご討論をいただきました。古瀬蔵 氏，内野一 氏（現在，NTT インテリジェントテクノロジ），高橋大和 氏，藤波進 氏（現在，NTT アドバンステクノロジ（株））には，テンプレート型と用例型の日英翻訳の研究に関して，ご協力とご討論をいただきました。秋葉泰弘 氏，春野雅彦 氏（ともに現在，（株）国際電気通信基礎技術研究所）には，日英翻訳の性能向上に関して，ご協力とご討論をいただきました。そのほか，様々な局面で研究にご協力いただいたNTT 研究所の方々に感謝します。データ分析や辞書構築にご協力いただいた小見佳恵 氏，上田洋美 氏，阿部さつき 氏，木村淳子 氏，渡邊いづみ 氏，井上浩子 氏，松尾三津恵 氏を始めとするNTT アドバンステクノロジ（株）の関係各位ほか，細井純子 氏，八木晶子 氏，ルール型システムの実現にご協力いただいた市井義健 氏，奥村信輔 氏，河村美砂子 氏を始めとするNTT ソフトウエア（株）の関係各位，テンプレート型システムの実現にご協力いただいた天井宏吉 氏を始めとする日本電子計算（株）の関係各位，言語分析ツールの実現にご協力いただいた赤坂哲治 氏を始めとするメリット（株）の関係各位，対訳データ収集にご協力いただいた鳴海武史 氏を始めとする（株）カナックの関係各位ほか，相澤弘 氏，武智しのぶ 氏，分部恵子 氏，森田千秋 氏、翻訳家の澤田信一 氏に深く感謝いたします。また，自然言語処理技術全般にいつも有益な示唆を与えてくださる山本和英 氏（長岡技術科学大学）と辞書構築技術の検討にご協力くださっているPaik Kyonghee 氏（（株）国際電気通信基礎技術研究所）に感謝します。

最後に，日ごろ研究を支えてくれる，妻はじめ家族，両親に感謝します。

付録 A　不完全な対訳データを利用する用例利用型翻訳

A.1 緒言

機械翻訳の方式として，ルール型翻訳と用例型翻訳の方法が提案され検討が続けられているが，いずれにも未解決の課題が多い。特に日英翻訳でば，言語類型が異なることもあり，実用に供せる段階とは言い難い[83]。本章では，ルール型翻訳および用例型翻訳の特徴を生かし，統計的手法を併用することにより，現状の技術レベルで構成可能な用例型翻訳の方式を検討する。検討に先立って，それぞれの翻訳方式の利点と欠点を概観する。

A.1.1 ルール型翻訳方式

現在市販されている機械翻訳ソフトの多くはこの方式によると思われる。いずれも，入力文を解析する処理（形態素解析，構文解析，意味解析など）と，構造変換を施した後の内部構造または中間言語から出力文を生成する処理を直列に配置した構成となっていると考えられる。各処理の動作は辞書とルールにより制御される。

翻訳システムの翻訳精度は各処理の精度の積で効いてくるため，辞書やルールを大規模化，高精度化する必要がある[47]。辞書やルールの規模が拡大すれば，それに伴って辞書項目やルール間の不整合が無視できなくなり，それらを整合させるには多大な工数を要するという問題がある。

また，解析のレベルを深くすると詳細な言語情報が使えるようになり，文の構造が明示的に捉えられるようになる反面，文全体の中での各要素の相互関係を喪失しがちになる。このため，要素合成的手法により訳文を生成すると，局所的には正しい訳出が行なわれたとしても，訳文全体として見たときには体裁が整っていないという問題が生じることもルール型の持つ問題と云える。

したがって，解析処理の直列構成に伴う解析精度の低下の回避と，要素合成的手法による訳文生成の回避が翻訳品質の向上を図る上での緊急の課題であると考えられる。

A.1.2 用例型翻訳方式

人間が翻訳を行なう際，既存の対訳を参考にして訳文を生成する過程を模することにより，類推により翻訳する手法[6]が提案された。これは，あらかじめ対訳例文集を用意しておき，翻訳対象文と類似した翻訳例を真似ることにより翻訳するもので，利用可能な翻訳例が見つかった場合には整った訳文が生成されるという利点がある。また，ルール型翻訳のように言語現象を個別に分析して辞書やルールを作る必要がなく，対訳例文を追加するだけで翻訳能力の向上が期待できるため，爆発的な研究の広がりを見せた。

なお，この方式では，適当な翻訳例が見つからなかった場合には翻訳結果が得られないということになる。しかし，これは必ずしも欠点とばかりは云えない。訳文が得られないことで，別の翻訳エンジンを起動するきっかけとして利用することができるからである。

用例型翻訳では，一般に，1 対 1 に対応付けられた大量の対訳コーパスの存在を前提とするほか，単語や句の対応付けや情報付与を要求されることが多い[84]。この条件が満たされれば，マニュアルの改版のような翻訳結果の再利用への適用は有効であるとされる。しかし，この用途は翻訳メモリという簡単な仕掛けによりすでに実現されている。

大量の対訳コーパスを入手すること自体が容易ではないが，仮にコーパスの量が確保されたとしても，首尾一貫した文分割や情報付与を行なうのも容易ではない。Kaji らは解析処理によりコーパスから翻訳テンプレートのセットを作成して利用する方法を提案しているが[85]，解析処理が修正されるたびに翻訳テンプレートを作成し直す必要があるという問題がある。

Cranias らは機能語と内容語の違いに着目したマッチング方法を提案しているが[86]，文分割と構成要素の対応付けの正しさを前提としているため，自動化は難しいと考えられる。

したがって，実用的な用例型翻訳を実現するには，言語間の対応関係と詳細な情報が付与された理想的な対訳用例集を構築することが困難であるという前提に立ったうえで，適切な用例を選択する方法を実現することが課題である。

A.1.3 融合型翻訳方式

融合型翻訳は，ルール型翻訳と用例型翻訳を併用し，そのいずれかの翻訳結果を選択して利用する方式である。この方式は，ルール型翻訳と用例型翻訳の双方の得失を引き継いでいるほか，最適な結果を選択するにはどうするかという新たな問題を生じるが，いずれか一方のエンジンだけでの翻訳するよりも精度の高い翻訳結果が得られることが期待できる。結果の選択方法としては，翻訳エンジンにあらかじめ優先順位をつけておく方法や，翻訳結果を統計的手法で点数付けして選択する方法が試みられている。

Brown は用例型翻訳をルール型翻訳と並行して走行させるシステム構成を提案している[87]。複合語の翻訳に用例型翻訳を適用しているに過ぎないが，限定的な適用であることがほかの言語への拡張も容易にしていると考えられる。

A.2 用例利用型翻訳方式の提案

ルール型翻訳の利点は解析処理により言語情報が得られること，用例型の利点は翻訳例の利用により整った訳文が生成される点にある。これらの利点を生かしながら，各処理の精度の積で全体の精度が決まる解析型翻訳の欠点と，対訳例文収集やタグ付与のコストがかかる用例型翻訳の欠点をカバーする方式を提案する。

まず，対訳用例を抽出する際には，文字レベルで入力文と類似する用例候補文を対訳データベースから抽出した後，双方に解析処理を適用することにより，得られた言語情報を用いて類似判定を行なう。解析処理としては，形態素解析，構文解析，意味解析などを必要に応じて使い分け，同形式異内容の表現が誤って選択されることを防止する。このように，類似判定の際に解析処理を適用することで，対訳データベースに言語情報タグを付与しておく事前処理の必要がなくなり，

解析処理の改変に伴うデータベースの再構成も不要となる。さらに，翻訳の際に，解析処理の精度への依存性を低減する効果も考えられる。すなわち，入力文および類似用例の双方に解析処理が適用されることから，同一の文字列に対しては同様の解析誤りが生じることになり，類似判定の際に解析誤りを相殺させることが可能となる。このため，1 文全体に解析処理を適用する場合よりも，実効的な解析精度は高くなると考えられる。

次に，対訳用例を決定する際には，単に入力文と最も類似する用例候補文に対応する対訳用例を選択するのではなく，抽出された対訳用例の訳文候補をチェックして，最も典型的な訳文を与える対訳用例であることを対訳用例の選択条件に加える。典型的でない対訳用例を破棄することにより，文脈に依存する対訳用例が選択される危険性を減少させることが可能となる。このため，対訳用例を収集する際，従来の用例型翻訳ほどは対訳用例の直訳性を要求されずに済むようになる。その一方で，対訳用例候補から典型的な対訳用例が得られることが，この方式に対する必要条件となる。

提案方式の概要は次の通りである。

(1)入力文 S_Iを解析し，文字レベルで類似する用例候補{ S_i }を抽出する。(6.2.2 節)
用例候補が得られなければ，ルール型翻訳により訳文を生成し，ステップ(4)に進む。
(2)入力文と最も整合する対訳用例{S_t,T_t}を選択する。(6.2.3 節)
(a)用例候補{ S_i }をそれぞれ解析し，入力文 S_Iとの類似度を計算する。
(b)用例候補{ S_i }に対応する訳文候補{ T_i }に対し，訳文 T_iの相互類似性に応じてグループ分割する。
(c)入力文 S_Iとの類似度，および，訳文グループの大きさと考慮して対訳用例{S_t,T_t}を選択する。
(3)得られた対訳用例{S_t,T_t}に基づいて訳文 T_Iを生成する。(6.2.4 節)
(a)入力文 S_Iと用例文 S_tの差分 d_iについて，用例訳文 T_tの対応箇所を調べる。
(b)差分 d_iをルール型翻訳で翻訳し，用例訳文 T_tの対応箇所と置換する。
(c)数の一致などにより訳文 T_Iを完成させる。
(4)訳文 T_Iを出力する。

A.3 用例利用型翻訳の適用例

A.3.1 対訳例文の収集

用例型翻訳の実用性を向上させるには大規模な対訳コーパスの実現性を議論しておくことがが必須である。前節で提案した方式は，翻訳を行なう際，抽出された用例候補文の訳文を類型化することにより先にチェックし，典型的な訳文を与える用例候補文を選択するようにしていることから，従来方式のような厳密な対応付けは必ずしも要求されないと考えられる。

大規模な対訳コーパスを構築するには，継続的にデータが得られる新聞記事の利用が考えら

れる。日本経済新聞の場合，日本語記事のCD-ROMのほかに，英語記事がテレコンデータベースに登録されており，記事単位で見たとき，半数の英文には直訳的に対応する日本文が存在することが知られている[88]。これらを利用することにより，緩い対応付け（厳密に対応させることはできないが，ある程度の直訳的な対応付けが可能な例文対）が行なわれた対訳データベースを構築することができる。

対訳データベースの構築には，対応する記事を発見することが必要であるが，Takahashiらの方法[89]により，自動的に対応関係を発見することができる。その後の文の対応付けには，Harunoらの方法[90]が利用できる。

対訳データベースには，TEI P3の文書定義[91]に従って，SGMLタグを付与する。ただし，これはデータベース化のために行なうタグ付与であり，従来手法で必要とされるような言語情報のタグ付与は行なわない。データベースの構成は，Lingua Project[92]を参考にして，次のように定める。対応情報を分離することにより，容易に対応付けを見直すことが可能となる。

a 記事単位，および，文単位で，ユニークなIDを付与する。
b 日本語記事と英語記事は，それぞれ別の領域に格納する。
c 日本語記事と英語記事の対応情報は，リンク領域に格納する。

また，対訳データベースの日本語には索引を付与し，例文検索の効率化を図る。索引としては，部分文字列を抑制したn-gramの手法[93]を用いて生成する。日本語の場合は文字単位に索引を生成するが，英語のように単語境界が明確な言語の場合は単語単位に生成を作成する。生成された索引は，Aoeらの方法[94]によりダブルアレイ型のtrie構造にしておくことにより，高速な検索を実現する。

A.3.2 用例候補文の抽出

入力文S_Iに対し，対訳データベースの索引を検索し，1個以上のn-gram文字列が一致するものを用例候補文として抽出する。

入力文として，次の文を例にとって説明する。

S_I: 日経平均10月物は続落。

この入力文に対し，2文字以上の文字列を作成し，対訳データベースの索引を検索する。この結果，「日経」，「平均」，「月物は続落」，「は続落」の4つのn-gram文字列に対して，次の3文が対訳候補文として対訳データベースから抽出される。

S_1: 日経平均9月物は続落。
S_2: 日経店頭平均は続落。
S_3: 8月物は続落。

このとき，対訳候補文が 1 文も抽出されなかった場合は，ルール型翻訳を用いることにより訳文が生成されるほか，次節以降の処理によりすべての対訳候補文が不適格とされた場合も，ルール型翻訳により訳文が生成される。

また，対訳候補文には同形式異内容の表現を持つ例文が含まれている可能性があるが，次節の訳文の類型化，または，入力文と用例候補文の解析情報を用いた類似判定により排除されることを期待する。

A.3.3 対訳用例の選択

6.3.3.1 用例候補文の類似度評価

類似度評価は例えば次のようにして行なう。

入力文 S_Iと，抽出された用例候補文 $\{S_1,S_2,S_3\}$ に対し，形態素解析[95]を適用し，単語分割を行なう。入力文 S_Iに含まれる単語に対し，先頭を 0 とし，以下，昇順に番号を付与する。用例候補文に対して，入力文 S_Iに含まれる単語についてはその番号を付与し，含まれない単語については十分大きい番号（下の説明では 99）を付与する。その上で，語順の一致の度合いに基づく類似度 M_0と，単語の一致度に基づく類似度 M_cを，それぞれ次のように定義する。

$$M_o(S_i, S_I) = \frac{S_i\text{のバブルソートに必要なスワップ回数}}{S_i\text{を反転するのに必要なスワップ回数}}$$

$$M_c(S_i, S_I) = \frac{S_I\text{と}S_i\text{の共通単語数}}{S_i\text{の単語数}}$$

この両者を次のように組み合わせることにより，入力文 S_Iと用例候補文 S_iの類似度 $M(S_i,S_I)$を次のように定義する[7]。ここで，入力文 S_Iと用例候補文 S_iの単語数に違いがある場合にペナルティを与えるため，相互に類似度を計算した平均値とする。

$$M(S_i, S_I) = ((1 - M_o(S_i, S_I)) \cdot M_c(S_i, S_I) + (1 - M_o(S_I, S_i)) \cdot M_c(S_I, S_i)) / 2$$

前節の例では，類似度は次のようになる。

S_I:	日経	平均	10	月	物	は	続落	。	
	0	1	2	3	4	5	6	7	
S_1:	日経	平均	9	月	物	は	続落	。	S_1 vs S_I: スワップ＝5，共通単語数=7
	0	1	99	3	4	5	6	7	S_I vs S_1: スワップ＝5，共通単語数=7
S_2:	日経	店頭	平均	は	続落	。			S_2 vs S_I: スワップ＝4，共通単語数=5
	0	99	1	5	6	7			S_I vs S_2: スワップ＝9，共通単語数=5

S_I:	日経	平均	10	月	物	は	続落	。
S_3:	8	月	物	は	続落	。		
	99	3	4	5	6	7		

S_3 vs S_I: スワップ＝5, 共通単語数=5

S_I vs S_3: スワップ＝15, 共通単語数=5

$M(S_1, S_I) = ((1-5/28)\cdot 7/8 + (1-5/28)\cdot 7/8)/2 = 0.719$

$M(S_2, S_I) = ((1-4/15)\cdot 5/6 + (1-9/28)\cdot 5/8)/2 = 0.518$

$M(S_3, S_I) = ((1-5/15)\cdot 5/6 + (1-15/28)\cdot 5/8)/2 = 0.423$

上記の例では, 類似判定に使用する言語情報を取得するための解析処理として形態素解析のみを使用している。必要に応じて, 構文解析や意味解析などの深い解析を適用することにより, 同形式異内容の用例候補文を排除することが可能となると考えられる。

例えば, 入力文が「私は電車に乗って学校へ行く。」で, 用例候補文が「半数は電車に乗って残りは歩いて行く。」である場合, 形態素レベルでは「に乗って」は一致しているが, 3.3 節で示したように構文構造が異なっている。形態素解析のみで類似判定を行なうと, この用例候補文は適格と判定される可能性がある。構文解析を適用することにより, 文の構造が異なっていることがわかるので, 用例候補文から排除することができる。

A.3.3.2 用例訳文のグループ化

用例候補文 $\{ S_i \}$ に対する用例訳文 $\{ T_i \}$ を, 相互の類似性に着目することによりグループ分けを行なう。

抽出された用例候補文 $\{ S_i \}$ は, 程度の差はあるにせよ, 入力文 S_I と類似しているはずであり, 相互に共通部分を持つ。用例候補文 $\{ S_i \}$ に対応する用例訳文 $\{ T_i \}$ も相互に共通性を持つことが期待される。そこで, 共通部分に着目して用例訳文をグループ分けすることにより, より多くの用例訳文が所属するグループを典型的な訳文であると推定する。このとき, 着目する共通部分が異なるのであれば, 1 つの用例訳文が複数のグループに所属することを妨げない。また, 用例候補文の中に入力文の表現と同形式異内容の表現を持つ文が取り出されている場合, それぞれの訳文は異なるグループに分類されることが期待される。

グループ化は, 例えば, 用例訳文 $\{ T_i \}$ の集合に含まれる各単語に対して, この集合内における出現頻度を集計し, この出現頻度に着目することにより行なう。ただし, 対象とする単語は内容語とし, 機能語は集計の対象とはしないこととする。

前節の例に対し, 次のような用例訳文が与えられるとする。(添字は出現頻度, 機能語は対象から除外。)

T_1: The Nikkei$_2$ average$_2$ September$_1$ contracts$_2$ were lower$_1$.

T_2: The Nikkei$_2$ over-the-counter$_1$ average$_2$ continued$_2$ declining$_2$.

T_3: August$_1$ contracts$_2$ continued$_2$ declining$_2$.

この例では, “continued declining”という連続する 2 単語が T_2 と T_3 に共通していることが分かる

ので，{ T_2, T_3 } が 1 つのグループとして推定され，残る { T_1 }が別のグループとされる。

ただし，“Nikkei”と“average”の 2 単語が共通することから，{ T_1, T_2 } を第 2 のグループとしても構わない。この場合，T_2 が 2 つのグループに重複して現れることになるが，このグループ分けは訳文の典型性の度合いを見極めるためのものであり，重複することには何らの問題はない。

A.3.3.3 対訳用例の選択

前節までに得られた評価値のうち，用例訳文のグループの大きさに基づいてグループの優先順位を設定し，さらにそのグループ内の用例候補文の類似度の高い順に優先順位を設定する。訳文の生成では，この優先順位に従って，利用可否を判定し，最初に利用可能な対訳用例を用いて訳文の生成が行なわれる。

前節の例では，用例訳文のグループの大きさに基づいて { T_2, T_3 } → { T_1 } の優先順位が設定され，さらに 6.3.3.1 節の用例候補文の類似度を用いることにより，T_2 → T_3 → T_1 の順に対訳用例を使用する優先順位が設定される。

A.3.4 対訳用例を用いた訳文生成

対訳用例を用いた訳文生成は次の手順で行なう。

第 1 優先の対訳用例を対象とし，入力文と用例原文の差分箇所を抽出する。入力文と用例原文の差分箇所を，それぞれ，対訳辞書検索，または，ルール型翻訳により翻訳する。用例原文の差分箇所の翻訳と用例訳文を比較し，一致する箇所を検出する。検出した一致箇所を，入力文の差分箇所の翻訳と置換し，数の一致などの補正を施すことにより，入力文の翻訳結果とする。

本章の方法をテンプレート翻訳と対比すると，用例原文の差分箇所とその翻訳結果が現れる用例訳文の該当箇所が，テンプレート翻訳ルールの変数部分に相当する。テンプレート翻訳ルールには変数に対する制約条件があらかじめ設定されているため，入力文に対するルールの適用可否が容易に判定される。これに対して，本章の方式では対訳用例に現れる表現のみが制約条件として使える情報であり，被覆性が極めて低い。しかし，訳文をグループ化したことにより，最も汎用的な訳文が得られる対訳用例が選択されているので，差分箇所を置換することによる訳文全体としての不整合の度合いは低いことが期待される。

不整合を低減させる目的で，入力文と用例原文の差分箇所が複合語の構成要素である場合，複合語全体を差分箇所として抽出するようにする。一般にルール型翻訳による複合語の翻訳精度は高いので，一部の構成要素を差し替えた時に危惧される局所的な不整合の発生を防止する。また，差分箇所が名詞句である場合は，用例原文の解析だけでなく，用例訳文の解析も行なったうえで対応関係を調べ，用例原文の名詞句に相当する用例訳文の該当箇所がひとまとまりでない場合は不整合が生じる恐れがあるため，いったんこの対訳用例の使用を保留する。しかし，他に適当な対訳用例が見出せない場合は，次善の策としてこの対訳用例を使用して翻訳する。

差分箇所に関して用例原文と用例訳文の対応が検出できなかった場合は，第 2 優先以降の対訳用例を対象として，翻訳結果の生成を試みる。すべての対訳用例に対して訳文生成が失敗す

る場合は，用例型翻訳を適用することができないということであるため，入力文に対してルール型翻訳により訳文を生成する。

前節の例では，翻訳は次のようにして生成される。

(1) 照合対象の対訳用例

前節で設定した優先順位に従い，対訳用例は S_2 T_2 を最初の照合対象とする。

S_1：日経平均 10 月物は続落。

S_2：日経店頭平均は続落。

T_2：The Nikkei over-the-counter average continued declining.

(2) 差分箇所の抽出

S_1：日経平均 10 月物は続落。

S_2：日経店頭平均は続落。

入力文 S_1 と用例原文 S_2 について，直接の差分となっている箇所は S_1 の「10 月物」と S_2 の「店頭」であるが，とも複合語の一部である。このため，S_1 の S_2 に対する差分箇所 d_1 は「日経平均 10 月物」，S_2 の S_1 に対する差分箇所 d_2 は「日経店頭平均」とする。

d_1（S_1 の S_2 に対する差分箇所）：日経平均 10 月物

d_2（S_2 の S_1 に対する差分箇所）：日経店頭平均

(3) 差分箇所の翻訳

入力文の用例原文に対する差分箇所 d_1 と用例原文の入力文に対する差分箇所 d_2 を，対訳辞書検索またはルール型翻訳によりそれぞれ翻訳する。d_1 はルール型翻訳により訳語が得られるが，d_2 は対訳辞書検索により訳語が得られる。

dt_1（S_1 の S_2 に対する差分箇所 d_1 の翻訳）：the Nikkei average October contracts

dt_2（S_2 の S_1 に対する差分箇所 d_2 の翻訳）：the Nikkei over-the-counter average

(4) 用例原文の差分箇所の訳と用例訳文の対応付け

用例原文 S_2 の入力文 S_1 に対する差分箇所 d_2 の翻訳である dt_2 を，用例訳文 T_2 の中に見出し，翻訳テンプレートを完成される。もし，差分箇所の翻訳が用例訳文に見出せない場合には，着目している対訳用例は不適格であるとして破棄し，次に優先順位が高い対訳用例を取り出し，入力文との照合を開始する。

dt_2：the Nikkei over-the-counter average

↓

T_2：The Nikkei over-the-counter average continued declining.

この例では T_2 内に dt_2 が見出せることから，この対訳用例は適格であると判定して，次節の入力文に対する訳文生成に進む。これは，対訳用例 S_2 T_2 から次のような翻訳テンプレートが生成されたことに相当する。

S_2：《変数 1》は続落。

T_2：《変数 1 の翻訳》continued declining.

(4) 入力文に対する訳文の生成と補正

用例訳文 T_2 を解析することにより，SVC 構造であること，S に相当する《変数 1 の翻訳》は単数であること，V に相当する “continued” は過去形であること，などの文法情報を把握しておく。次に，《変数 1 の翻訳》に入力文の差分箇所の翻訳 dt_I を代入し，翻訳結果 T_I を得るとともに，英語の文法情報に従って訳文の補正を行なう。この例では動詞 “continued” は過去形であるため特段の補正は必要としない。以上により，次のような翻訳結果が得られる。

dt_I：the Nikkei average October contracts

↓

T_I：The Nikkei Average October contracts continued declining.

A.4 結言

用例を利用する翻訳の利点は，慣用的な表現や分野に依存する表現に対しても，システム構成を変更しないで適用できるところにある。また，対訳用例は，逆方向の翻訳にも使用できることから，システムの拡張性も高いと考えられる。

従来の用例型翻訳では，用例を追加することによりシステム性能を不断に向上させることができるという定性的な特徴が主張されるだけで，詳細な対応付け情報が整備された大規模対訳コーパスの存在を前提としてシステム提案が行なわれ，対訳コーパスの構築についてはあまり議論されていなかった。このため，実用性という観点では多くの課題が残されたままとなっている。

これに対し，本章では，対訳コーパスの構築可能性を検討し，新聞記事を利用することにより，緩い対応付けがされた対訳コーパスであれば実現可能であること，複数の対訳用例を相互に比較することにより適切な用例を自動的に選択できる可能性が高いこと，翻訳時に入力文と対訳用例を解析して言語情報を利用するようにすればコーパスへの自然の情報付与が不要となること，また，解析誤りによる翻訳への悪影響を極小化できることを示し，新しい用例利用型翻訳の方式提案を行なった。

参考文献

[1] 栗原俊彦 (1973). 自然言語の機械処理. 情報処理, Vol.14, No.4, pp.267-281.

[2] 首藤公昭 (1974). 専門分野を対象とした日英機械翻訳について. 情報処理, Vol.14, No.9, pp.661-668.

[3] 坂井, 杉田 (1966). 機械による英和翻訳. 電子通信学会論文誌, Vol.49, No.2.

[4] 長尾, 辻井, 矢田, 柿元 (1982). 科学技術論文表題の英和翻訳システム. 情報処理学会論文誌, Vol.23, No.2, pp.202-210.

[5] *Proceedings of TMI-92* (1992).

[6] Makoto Nagao (1984). A framework of mechanical translation between Japanese and English by analogy principle. *Elithorn and Banerji (eds.), Artificial and Human Intelligence*, pp.179-180.

[7] Satoshi Sato (1992). CTM. An example based translation aid system. In *Proceedings of COLING-92* , pp. 1259-1263.

[8] P. F. Brown, C. John, S. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercar & P. S. Roossin (1990). A statistical approach to machine translation. *Computational Linguistics*, Vol.16, No.2, pp.79-85.

[9] T. Watanabe & E. Sumita (2002). Bidirectional decoding for statistical machine translation. In *Proceedings of COLING-2002*, pp. 1075-1085.

[10] 森, 山地 (1997). 日本語の情報量の上限の推定. 情報処理学会論文誌, Vol.38, No.11, pp.2191-2199.

[11] 三浦つとむ (1967). 認識と言語の理論（第 1-3 巻）. 勁草書房.

[12] 水谷, 石綿, 荻野, 賀来, 草薙 (1983). 文法と意味 *I*（朝倉日本語新講座 3）. 朝倉書店.

[13] 宮崎正弘 (1986). 日本文音声出力のための言語処理に関する研究. 東京工業大学学位論文.

[14] 南不二男 (1974). 現代日本語の構造. 大修館書店.

[15] M. Nagao & S. Mori (1994). A New Method of *N*-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese. In *Proceedings of COLING-94*, pp.611-615.

[16] Makoto Nagao (1993). Varieties of Heuristics in Sentence Parsing. *Invited Talk at International Workshop on Parsing Technology*.

[17] 黒橋, 長尾 (1992). 長い日本語文における並列構造の推定. 情報処理学会論文誌, Vol.33, No.8, pp.1022-1031.

[18] 黒橋, 長尾 (1994). 並列構造の検出に基づく長い日本語文の構文解析. 自然言語処理,

Vol.1, No.1, pp.35-58.
[19] 長尾真 編著 (1996). 自然言語処理. 岩波書店.
[20] 京都大学大学院情報学研究科言語メディア研究室. 京都大学テキストコーパス Version 3.0. http://www.kc.t.u-tokyo.ac.jp/nl-resource/corpus.html.
[21] 工藤, 松本 (2002). チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol.43, No.6, pp.1834-1842.
[22] 奥村, 池野, 松下, 山本, 永田 (1993). 日本語文の並列構造を利用した長文解析方式. 第 *7* 回人工知能学会全国大会論文集, 17-4.
[23] 白井, 横尾, 木村, 小見 (1993). 日本語従属節の依存構造に着目した係り受け解析. 第 *47* 回情報処理学会全国大会論文集, 3M-1.
[24] 池野, 奥村, 松下, 山本, 永田 (1993). 日本語長文の翻訳における副詞呼応範囲の優先構造化方式. 第 *7* 回人工知能学会全国大会論文集, 17-6.
[25] 林, 奥, 石崎 (1986). 日英翻訳システム ALT-J/E における日英変換枝術. 第 *33* 回情報処理学会全国大会論文集, 6J-3.
[26] 林良彦 (1987). 結合価構造に基づく日本語解析. 情報処理学会研究報告, 87-NL-62.
[27] 横尾, 林 (1987). 日本語埋め込み構造の解析. 第 *1* 回人工知能学会全国大会論文集, 7-2.
[28] 白井諭 (1987). 日英翻訳システム ALT-J/E におけるテーブル駆動型日本語文節間係り受け解析法. 第 *34* 回情報処理学会全国大会論文集, 5W-5.
[29] 山田孝雄 (1936). 日本文法学概論. 宝文館.
[30] 時枝誠記 (1941). 国語学原論. 岩波書店.
[31] 渡辺実 (1953). 叙述と陳述 －述語文節の構造－, 国語学, 13/14.
[32] 芳賀綏 (1954). “陳述”とは何もの?. 国語国文, Vol.23, No.4.
[33] 服部四郎 (1957). ソスュールの langue と言語過程説. 言語学の方法, 岩波書店.
[34] 林四郎 (1960). 基本文型の研究. 明治書院.
[35] 南不二男 (1964). 述語文の構造. 日本の言語学, 服部編, 大修館書店.
[36] 南不二男 (1964). 複文. 講座現代語 *6*, 時枝, 遠藤監修, 明治書院.
[37] 三浦つとむ (1975). 日本語の文法. 勁草書房.
[38] J. Carbonell, et al. (1992). *JTEC Panel Report on “Machine Translation in Japan”*. Coordinated by Loyola College in Maryland.
[39] M. Rimon, M. McCord, U. Schwall & P. Martinez (1991). Advances in Machine Translation Research in IBM. In *Proceedings of MT SUMMIT III*, pp.11-18.
[40] *Proceedings of COLING '92* (1992).
[41] 池原, 白井 (1990). 日英機械翻訳機能試験項目の体系化. 電子情報通信学会技術研究報告, NLC90-43, pp.17-24.

[42] Satoru Ikehara (1992). Criteria for Evaluating the Linguistic Quality of Japanese to English MT System. In *Proceedings of MT Evaluation Workshop*, pp.58-59.

[43] 池原, 宮崎, 白井, 林 (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.

[44] Satoru Ikehara (1989). Multi-Level Machine Translation Method. *Future Computer Systems*, Vol.2, No.3, pp.261-274.

[45] S. C. Chen, J. N. Wang, J. S. Chang & K. Y. Su (1991). ArchTran: A Corpus-based Statistics-oriented English Chinese Machine Translation System. In Proceedings *of MT SUMMIT III*, pp.11-18.

[46] Sergei Nirenburg (1989). KBMT-89-A Knowledge Based MT Project at Carnegie Melon University. In *Proceedings of MT SUMMIT II*, pp.141-147.

[47] 池原, 宮崎, 横尾 (1993). 日英機械翻訳のための意味解析用の知識とその分解能. 情報処理学会論文誌, Vol.34, No.8, pp.1692-1704.

[48] O. Furuse & H. Iida (1992). Cooperation between Transfer and Analysis in Example-Based Framework, In *Proceedings of COLING-92*, pp.645-651.

[49] Makoto Nagao (1992). Some Rationales and Methodologies for Example-based Approach. In *Proceedings of Workshop on Future Generation Natural Language Processing*, Manchester.

[50] 長尾真 (1985). 制限言語の提案. 自然言語処理ジンポジウム, 情報処理学会.

[51] 長尾, 田中, 辻井 (1984). 制御言語にもとづく文章作成援助システム. 情報処理学会研究報告, NL-44-5.

[52] 長尾真 (1983). 科技庁機械翻訳プロジェクトの概要. 情報処理学会研究報告, NL-38-2.

[53] 辻井, 長尾 (1985). 日英翻訳過程での処理とその翻訳結果への反映. 情報処理学会研究報告, NL-47-10.

[54] 白井諭 (1990). 日本文自動書き替えによる構文多義の解消. 第 *41* 回情報処理学会全国大会論文集, 4S-6.

[55] 池原, 安田, 島崎, 高木 (1987). 日本文訂正支援システム（REVISE）. 研究実用化報告, Vol.36, No.9, p.1159-1167.

[56] H. Nakaiwa & S. Ikehara (1992). Zero Pronoun Resolution in a Japanese to English Machine Translation System using Verbal Semantic Attributes. In Proceedings *of the 3rd Conference on Applied Natural Language Processing*, pp.201-208.

[57] Automatic Language Processing Advisory Committee (1966). *Language and Machines: Computers in Translation and Linguistics*. Division of Behavioral Sciences, National Academy of Science, National Research Council Publication 1416.

[58] 白井, 池原, 阿部, 松尾 (1994). 日本文書き替え処理における制御ルールの類型情報の抽出.

情報処理学会第 *49* 回全国大会, 4G-12, Vol.3, pp.243-244.

[59] 白井, 池原, 松尾, 兵藤 (1994). 日本文書き替え処理における制御機能の構成について. 情報処理学会第 *49* 回全国大会, 4G-13, Vol.3, pp.245-246.

[60] 黒橋, 長尾 (1992). 格フレーム選択における意味マーカと例文の有効性について. 情報処理学会研究報告, NL-91-11.

[61] H. Almuallim, Y. Akiba, T. Yamazaki, A. Yokoo & S. Kaneda (1994). A tool for the acquisition of Japanese to English machine translation rules using inductive learning techniques. In *Proceedings of CAIA94*, pp.194-201.

[62] H. Almuallim, Y. Akiba, T. Yamazaki, A. Yokoo & S. Kaneda (1994). Two methods for learning ALT-J/E translation rules from examples and a semantic hierarchy. In *Proceedings of COLING -94*, pp.57-63.

[63] 金田, 秋葉, 石井, アルムアリム (1994). 事例に基づく英語動詞選択ルールの修正型学習方式.「自然言語処理における学習」シンポジウム論文集, pp.158-165.

[64] 横尾, 中岩, 白井, 池原 (1994). 日英機械翻訳用スケルトンーフレッシュ型構文意味辞書の構成. 第 *48* 回情報処理学会全国大会論文集, 6Q-8, Vol.3, pp.139-140.

[65] 小島, 竹林 (1984). ライトハウス和英辞典, 第1版. 研究社.

[66] 情報処理振興事業協会 技術センター（編）(1987). 計算機用日本語基本動詞辞書 *IPAL (Basic Verbs)*, 解説編&辞書編.

[67] 白井, 兵藤, 上田, 横尾, 池原(1995). 日英機械翻訳用構文意味辞書の作成支援, *H7* 年電気関係学会関西支部連合大会, G14-3, p.G368.

[68] 林巨樹（編）(1985). 現代国語例解辞典, 第一版. 小学館.

[69] 林巨樹（編）(1997). 現代国語例解辞典, 第二版. 小学館.

[70] 情報処理振興事業協会 技術センター（編）(1990). 計算機用日本語基本形容詞辞書 *IPAL (Basic Adjectives)*, 解説編&辞書編.

[71] K. W. Church & P. Hanks (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, Vol.16, No.1, pp.22-29.

[72] F. A. Smadja & K. R. MeKeown (1990). Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp.252-259.

[73] 北, 小倉, 森元, 矢野 (1993). 仕事量基準を用いたコーパスからの定型表現の自動抽出. 情報処理学会論文誌, Vol.34, No.9, pp.1937-1943.

[74] K. Kita, Y. Kato, T. Omoto & Y. Yano (1994). A comparative study of automatic extraction of collocations from corpora: mutual information vs. cost criteria. *Journal of Natural Language Processing*, Vol.1, No.1, pp.21-33.

[75] F. Smadja (1993). Retrieving collocations from text. *Computational Linguistics*, Vol.19, No.9, pp.143-177.

[76] 浦谷則好 (1995). ニュース原稿データベースからの表現パターンの抽出. 第 50 回情報処理学会全国大会論文集, 1R-8.

[77] 新納, 井佐原 (1995). 疑似 N グラムを用いた助詞的定型表現の自動抽出. 情報処理学会論文誌, Vol.36, No.1, pp.32-40.

[78] R. Collier (1994). *N*-gram cluster identification during empirical knowledge representation generation, In *Proceedings of COLING-94*, pp.1054-1058.

[79] 加藤, 相沢 (1993). 外電ニュースの定型文抽出とその英日機械翻訳. 情報処理学会研究報告, NL-93-2 .

[80] 内野, 池原, 白井 (1996). 弱抑制による連鎖共起表現の抽出とそれに基づく離散共起表現の抽出. 言語処理学会第 2 回年次大会, B6-4, pp.257-260.

[81] 例えば, 近藤泰弘,近藤みゆき(2001). 平安時代古典語古典文学研究のための N-gram を用いた解析手法. 言語処理学会第 7 回年次大会, C3-4.

[82] O. Sander, I. Fischer & H. Kirsch (2002). Extracting collocations from syntactically annotated corpora. In *Proceedings of FGML2002*, pp. 127-134,

[83] 成田一 (1996). 言語類型と機械翻訳. 情報処理学会研究報告, 96-NL-114-21, pp.143-150.

[84] Victor Sadler (1989). *Working with Analogical Semantics: Disambiguation techniques in DLT*. FORIS Publications.

[85] H. Kaji, Y. Kida & Y. Morimoto (1992). Learning translation templates from bilingual text. In *Proceedings of COLlNG-92* , pp. 672-678.

[86] L. Cranias, H. Papageorgiou & S. Piperidis(1995). A matching technique in example-based machine translation.

[87] Ralf D. Brown (1996). Example-based machine translation in the pangloss system. In *Proceedings of COLING-96* , pp. 125-130.

[88] 白井, 松尾, 瀬下, 藤波, 池原 (1995). 新聞記事日英対訳コーパスの構築(3), －記事の特徴分析と文の対応関係の検討－. 平成 7 年度(第 48 回)電気関係学会九州支部連合大会, p.857.

[89] Y. Takahashi, S. Shirai & F. Bond (1997). A method of automatically aligning Japanese and English newspaper articles. In *Proceedings of NLPRS-97,* pp.49-54.

[90] M. Haruno & T. Yamazaki (1996). High-performance bilingual text alignment using statistical and dictionary information. In *34th Annual Conference of the Association for Computational Linguistics* , pp. 131-138.

[91] C. M. Sperberg-McQueen & Lou Burnard, eds. (1994). *Guidelines for Electronic Text*

Encoding and Interchange . Chicago, Oxford.

[92] P. Bonhomme & L. Romary (1995). The lingua parallel concordancing project. Managing multilingual texts for educational purpose. In *Proceedings of Language Engineering '95*.

[93] S. Ikehara, S. Shirai & H. Uchino (1996). A statistical method for extracting uninterrupted and interrupted collocations from very large corpora. In *Proceedings of COLING-96* , pp. 574-579.

[94] J. Aoe, K. Morimoto & T. Sato (1992). An efficient implementation of trie structures. *Software Practice & Experiments* , 22(9), pp. 695-721.

[95] S. Ikehara, S. Shirai, A. Yokoo & H. Nakaiwa (1991). Toward an MT system without pre-editing - effects of new methods in ALT-J/E-. In *Proceedings of MT Summit III* , pp. 101-106.

執筆論文リスト

［論文誌］

1. 池原悟, 白井諭 (1984.3). 単語解析プログラムによる日本文誤字の自動検出と二次マルコフモデルによる訂正候補の抽出. 情報処理学会論文誌, Vol.25, No.2, pp.298-305.
2. 池原悟, 宮崎正弘, 白井諭, 林良彦 (1987.12). 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.
3. 池原悟, 宮崎正弘, 白井諭, 横尾昭男 (1992.11). **An evaluation method for MT systems and its application to ALT-J/E**. 人工知能学会誌, Vol.7, No.6, pp.1077-1086.
4. 池原悟, 白井諭, 小倉健太郎 (1994.7). 言語表現体系の違いに着目した日英機械翻訳機能試験項目の構成. 人工知能学会誌, Vol.9, No.4, pp.93-103.
5. 池原悟, 白井諭, 横尾昭男, Francis Bond, 小見佳恵 (1995.1). 日英機械翻訳における利用者登録語の意味属性の自動推定. 自然言語処理, Vol.2, No.1, pp.3-17.
6. 白井諭, 池原悟, 河岡司, 中村行宏 (1995.1). 日英機械翻訳における原文自動書き替え型翻訳方式とその効果. 情報処理学会論文誌, Vol.36, No.1, pp.12-21.
7. 宮崎正弘, 白井諭, 池原悟 (1995.7). 言語過程説に基づく日本語品詞の体系化とその効用. 自然言語処理, Vol.2, No.3, pp.3-25.
8. 白井諭, 池原悟, 横尾昭男, 木村淳子 (1995.10). 階層的認識構造に着目した日本語従属節間の係り受け解析の方法とその精度. 情報処理学会論文誌, Vol.36, No.10, pp.2353-2361.
9. 池原悟, 白井諭, 河岡司 (1995.11). 大規模日本語コーパスからの連鎖型および離散型共起表現の自動抽出法. 情報処理学会論文誌, Vol.36, No.11, pp.2584-2596.
10. 中岩浩巳, 白井諭, 池原悟 (1997.11). 日英機械翻訳における語用論的・意味論的制約を用いたゼロ代名詞の文章外照応解析. 情報処理学会論文誌, Vol.38, No.11, pp.2167-2178.
11. 春野雅彦, 白井諭, 大山芳史 (1998.12). 決定木を用いた日本語係受け解析. 情報処理学会論文誌, Vol.39, No.12, pp.3177-3186.
12. Masahiko Haruno, Satoshi Shirai & Yoshifumi Ooyama (1999.2). **Using decision trees to construct a practical parser**. *Machine Learning*, Vol.34, Nos.1/2/3, pp.131-149.
13. 内野一, 白井諭, 横尾昭男, 大山芳史, 古瀬蔵 (2001.6). 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.1168-1174.
14. Yasuhiro Akiba, Hiromi Nakaiwa, Yoshifumi Ooyama & Satoshi Shirai (2001.12). **Interactive generalization of a translation example using queries based on a semantic hierarchy**. *International Journal on Artificial Intelligence Tools*, Vol.10, No.4, pp.675-690.

15. 畑山満美子, 松尾義博, 白井諭 (2002.7). 重要語句抽出による新聞記事自動要約. 自然言語処理, Vol.9, No.4, pp.55-73.
16. Kyoughee Paik, Hiromi Nakaiwa & Satoshi Shirai (2004.1). **Direct machine translation of Japanese to Korean**. *Harvard Studies in Korean Linguistics X*, pp.159-172.

［国際会議 ］

1. Masahiro Miyazaki, Shigeki Goto, Yoshifumi Ooyama & Satoshi Shirai (1983.6.17-19). **Linguistic processing in a Japanese-text-to-speech-system**. In *Proceedings of ICTP '83* (Tokyo), pp.315-320.
2. Satoru Ikehara, Satoshi Shirai, Akio Yokoo & Hiromi Nakaiwa (1991.7.1-4). **Toward an MT system without pre-editing, －Effects of new methods in ALT-J/E－**. In *Proceedings of MT SUMMIT III* (Washington, D.C., USA), pp.101-106.
3. Satoshi Shirai, Satoru Ikehara & Tsukasa Kawaoka (1993.7.14-16). **Effects of automatic rewriting of source language within a Japanese to English MT system**. In *Proceedings of TMI '93* (Kyoto, Japan), pp.226-239.
4. Kentaro Ogura, Akio Yokoo, Satoshi Shirai & Satoru Ikehara (1993.10.16-22). **Japanese to English machine translation and dictionaries**. In *Proceedings of 44th Congress of the International Astronautical Federation* (Graz, Austria).
5. Satoru Ikehara, Satoshi Shirai, Kentaro Ogura, Akio Yokoo, Hiromi Nakaiwa & Tsukasa Kawaoka (1994.8.28-9.2). **ALT-J/E, a Japanese to English machine translation system for communication with translation**. In *Proceedings of IFIP 13th World Computer Congress* (Humburg, Germany), Vol.2, pp.80-85.
6. Yoshihiro Matsuo, Satoshi Shirai, Akio Yokoo & Satoru Ikehara (1994.9.14-16). **Direct parse tree translation in cooperation with the transfer method**. In *Proceedings of NeMLaP* (Manchester, UK), pp.144-149.
7. Satoru Ikehara, Satoshi Shirai, Akio Yokoo, Francis Bond & Yoshie Omi (1994.10.13-15). **Automatic aquisition of semantic attributes for user defined words in Japanese to English machine translation**. In *Proceedings of ANLP '94* (Stuttgart, Germany), pp.184-185.
8. Hiromi Nakaiwa, Satoshi Shirai, Satoru Ikehara & Tsukasa Kawaoka (1995.3.27-29). **Extrasentential resolution of Japanese zero pronouns using semantic and pragmatic constraints**. In *Proceedings of AAAI '95 Spring Symposium* (San Francisco, USA), pp.99-105.
9. Satoru Ikehara, Satoshi Shirai, Kentaro Ogura, Akio Yokoo, Hiromi Nakaiwa & Tsukasa Kawaoka (1995.10.3-11). **Multi-level machine translation method for communication with**

translation. In *Proceedings of World TELECOM 95, Technology Summit* (Geneva, Switzerland), Vol.2, pp.623-627.

10. Yoshihiro Matsuo, Satoshi Shirai & Satoru Ikehara (1995.12.4-7). **Changing syntactic classes in transfer-based machine translation**. In *Proceedings of NLPRS '95* (Seoul, Korea), Vol.1, pp.432-437.
11. Satoshi Shirai, Satoru Ikehara, Akio Yokoo & Hiroko Inoue (1995.12.4-7). **The quantity of valency pattern pairs required for Japanese to English MT and their compilation**. In *Proceedings of NLPRS '95* (Seoul, Korea), Vol.1, pp.443-448.
12. Satoru Ikehara, Satoshi Shirai & Hajime Uchino (1996.8.5-9). **A statistical method for extracting uninterrupted and interrupted collocations from very large corpora**. In *Proceedings of COLING-96* (Copenhagen, Denmark), Vol.1, pp.574-579.
13. Hiromi Nakaiwa & Satoshi Shirai (1996.8.5-9). **Anaphora resolution of Japanese zero pronouns with deictic reference**. In *Proceedings of COLING-96* (Copenhagen, Denmark), Vol.2, pp.812-817.
14. Satoru Ikehara, Satoshi Shirai & Francis Bond (1996.8.12-14). **Approaches to disambiguation in ALT-J/E**. In *Procedings of MIDDIM-96* (Grenoble, France), pp.107-117.
15. Kentaro Ogura, Satoshi Shirai & Francis Bond (1997.7.23-25). **English adverb processing in Japanese-to-English machine translation**. In *Proceedings of TMI-97* (Santa Fe, USA), pp.95-102.
16. 小倉健太郎, 中岩浩巳, 横尾昭男, 白井諭, 宮崎正弘, 池原悟 (1997.8.27-28). 日英機械翻訳とシソーラス. 第 *5* 回国立国語研究所国際シンポジウム (Tokyo, Japan), pp.154-161.
17. Satoshi Shirai, Francis Bond & Yamato Takahashi (1997.12.2-4). **A hybrid rule and example-based method for machine translation**. In *Proceedings of NLPRS'97* (Phuket, Thailand), pp.49-54.
18. Yamato Takahashi, Satoshi Shirai & Francis Bond (1997.12.2-4). **A method of automatically aligning Japanese & English newspaper articles**. In *Proceedings of NLPRS'97* (Phuket, Thailand), pp.657-660.
19. Francis Bond & Satoshi Shirai (1997.12.5). **Practical and efficient organization of a large valency dictionary**. In *Proceedings of NLPRS'97 Multilingual Workshop* (Phuket, Thailand).
20. Satoshi Shirai, Satoru Ikehara, Akio Yokoo & Yoshifumi Ooyama (1998.5.21-22). **Automatic rewriting method for internal expressions in Japanese to English MT and its effects**. In *Proceedings of CLAW'98* (Pittsburgh, USA), pp.62-75.
21. Francis Bond, Daniela Kurz & Satoshi Shirai (1998.8.10-14). **Anchoring floating quantifiers in Japanese-to-English machine translation**. In *Proceedings of COLING-ACL'98* (Montreal,

Canada), pp.152-159.

22. Masahiko Haruno, Satoshi Shirai & Yoshifumi Ooyama (1998.8.10-14). **Using decision trees to construct a practical parser**. In *Proceedings of COLING-ACL'98* (Montreal, Canada), pp.505-511.
23. Eiichiro Sumita, Setsuo Yamada, Kazuhide Yamamoto, Michael Paul, Hideki Kashioka, Kai Ishikawa & Satoshi Shirai (1999.9.13-17). **Solutions to problems inherent in spoken-language translation: the ATR-MATRIX approach**. In *Proceedings of MT-SUMMIT VII* (Singapore), pp.229-235.
24. Hideki Kashioka & Satoshi Shirai (2000.5.31-6.2). **Automatically expansion of thesaurus entries with a different thesaurus**. In *Proceedings of LREC'2000* (Athens, Greece), pp.363-366.
25. Yves Lepage, Nicolas Auclerc & Satoshi Shirai (2000.10.16-20). **A tool to build a treebank for conversational Chinese**. In *Proceedings of ICSLP 2000* (Beijing, China), pp.985-988.
26. Yasuhiro Akiba, Hiromi Nakaiwa, Satoshi Shirai & Yoshifumi Ooyama (2000.11.13-15). **Interactive generalization of a translation example using queries based on a semantic hierarchy**. In *Proceedings of ICTAI00* (Vancouver, Canada), pp.326-332.
27. Satoshi Shirai, Kazuhide Yamamoto & Kazutaka Takao (2001.5.14-16). **Construction of a dictionary for translating Japanese phrases into one English word**. In *Proceedings of ICCPOL 2001* (Seoul, Korea), pp.3-8.
28. Kazuhide Yamamoto, Satoshi Shirai, Masashi Sakamoto & Yujie Zhang (2001.5.14-16). **Sandglass: Twin paraphrasing spoken language translation**. In *Proceedings of ICCPOL 2001* (Seoul, Korea), pp.154-159.
29. Satoshi Shirai & Kazuhide Yamamoto (2001.5.14-16). **Linking English words in two bilingual dictionaries to generate another language pair dictionary**. In *Proceedings of ICCPOL 2001* (Seoul, Korea), pp.174-179.
30. Toshiyuki Takezawa, Satoshi Shirai & Yoshifumi Ooyama (2001.5.14-16). **Characteristics of colloquial expressions in a bilingual travel conversation corpus**. In *Proceedings of ICCPOL 2001* (Seoul, Korea), pp.384-389.
31. Satoshi Shirai, Kazuhide Yamamoto & Kyonghee Paik (2001.8.22-24). **Overlapping constraints of two step selection to generate a transfer dictionary**. In *Proceedings of ICSP 2001* (Taejon, Korea), Vol.2, pp.731-736.
32. Kyonghee Paik & Satoshi Shirai (2001.8.22-24). **Exploiting linguistic similarities for machine translation: a case study of Japanese-to-Korean**. In *Proceedings of ICSP 2001* (Taejon, Korea), Vol.2, pp.737-742.

33. Nicholas Auclerc, Yves Lepage & Satoshi Shirai (2001.9.10-14). **Case study: porting an NLP application to Unicode**. In *Proceedings of IUC19* (San Jose, USA), Part.3, B201:1-22.
34. Chenqing Zong, Yujie Zhang, Kazuhide Yamamoto, Masashi Sakamoto & Satoshi Shirai (2001.11.27-29). **Paraphrasing Chinese utterances in spoken language translation system**. (in Chinese) In *Proceedings of ICCC 2001* (Singapore), pp.395-401.
35. Chenqing Zong, Yujie Zhang, Kazuhide Yamamoto, Masashi Sakamoto & Satoshi Shirai (2001.11.27-29). **Approach to spoken Chinese paraphrasing based on feature extraction**. In *Proceedings of NLPRS-2001* (Tokyo), pp.551-556.
36. Mamiko Hatayama, Yoshihiro Matsuo & Satoshi Shirai (2001.11.27-29). **Summarizing newspaper articles using extracted informative and functional words**. In *Proceedings of NLPRS-2001* (Tokyo), pp.593-600.
37. Satoshi Shirai, Kazuhide Yamamoto & Francis Bond (2001.11.30). **Japanese-English paraphrase corpus**. In *Proceedings of Workshop on Language Resources in Asia, NLPRS-2001* (Tokyo), pp.23-30.
38. Kyonghee Paik, Francis Bond & Satoshi Shirai (2001.11.30). **Using multiple pivots to align Korean and Japanese lexical resources**. In *Proceedings of Workshop on Language Resource in Asia, NLPRS-2001* (Tokyo), pp.63-70.
39. Satoshi Shirai, Kazuhide Yamamoto, Francis Bond & Hozumi Tanaka (2002.5.29-31). **Towards a thesaurus of predicates**. In *Proceedings of LREC 2002* (Canary Islands, Spain), Vol.6, pp.1965-1972.
40. Kyoughee Paik, Hiromi Nakaiwa & Satoshi Shirai (2003.7.11-13). **Direct machine translation of Japanese to Korean**. In *Proceedings of Harvard Biennial International Symposium on Korean Linguistics* (Cambridge, USA), pp.96-98.
41. Kyonghee Paik, Satoshi Shirai & Hiromi Nakaiwa (2004.8.28). **Automatic construction of a transfer dictionary considering directionality**. In *Proceedings of MLR 2004, COLING-2004*, (Geneva, Switzerland), pp.31-38.

［NTT 技術誌 ］

1. 池原悟, 中園薫, 白井諭 (1987). キーワード自動抽出システム**(INDEXER)**. 研究実用化報告, Vol.36, No.9, pp.1151-1158.

2. Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai & Akio Yokoo (1989). **An approach to machine translation method based on constructive process theory**. *Review of the Electrical Communications Laboratories*,Vol.37, No.1, pp.39-44.
3. 白井諭, 池原悟, 横尾昭男, 中岩浩巳 (1991). 前編集不要の日英機械翻訳の実現に向けて. *NTT R&D*, Vol.40, No.7, pp.897-904.
4. 大山芳史, 白井諭, 横尾昭男, 藤波進 (1997). 日英機械翻訳技術と市況速報への適用. *NTT* 技術ジャーナル, Vol.9, No.6, pp.73-76.
5. 八巻俊文, 大山芳史, 白井諭, 横尾昭男 (1997). 日英機械翻訳システム **ALT-J/E** の研究開発. *NTT R&D*, Vol.46, No.12, pp.1391-1398.
6. 白井諭, 横尾昭男, 松尾義博, 大山芳史 (1997). 日英翻訳のための日本語解析技術. *NTT R&D*, Vol.46, No.12, pp.1399-1404.
7. 白井諭, 横尾昭男, 内野一, 松尾義博 (1997). 日英変換技術と意味辞書. *NTT R&D*, Vol.46, No.12, pp.1405-1410.
8. 内野一, 春野雅彦, 高橋大和, 白井諭 (1997). 機械翻訳辞書構築支援ツール. *NTT R&D*, Vol.46, No.12, pp.1425-1432.

［研究会・シンポジウム等 ］

1. Hidehiko Sanada, Satoshi Shirai, Hikaru Nakanishi & Yoshikazu Tezuka (1980.10). **On job allocation algorithms in distributed processing netwoks**. *Technology Reports of The Osaka University*, Vol.30, No.1567, pp.457-462.
2. 白井諭, 井上健, 中西暉, 真田英彦, 手塚慶一 (1979.8.24). 分散処理網における処理ホスト決定アルゴリズムと網トポロジについて. 電子通信学会技術研究報告, SE79-64, pp.73-80.
3. 宮崎正弘, 白井諭, 大山芳史, 後藤滋樹, 池原悟 (1983.6.16-17). 日本文音声出力のための言語処理. 情報処理学会「自然言語処理技術」シンポジウム, pp.5-16.
4. 中園薫, 白井諭 (1984.11.6-7). 日本語索引自動生成システム. 情報処理学会「自然言語処理技術」シンポジウム, pp.19-25.
5. 池原悟, 白井諭 (1990.12.21). 日英機械翻訳機能試験項目の体系化. 電子情報通信学会技術研究報告, NLC90-43, pp.17-24.
6. 宮崎正弘, 池原悟, 白井諭 (1992.1.7-8). 言語の過程的構造と自然言語処理.「自然言語処理の新しい応用」シンポジウム, pp.60-69.
7. 池原悟, 宮崎正弘, 白井諭 (1992.1.7-8). 言語過程説から見た多段翻訳方式の意義.「自然言語処理の新しい応用」シンポジウム, pp.139-140.
8. 白井諭, 宮崎正弘, 池原悟 (1992.1.7-8). 言語過程説から見た日本語述語の構造.「自然言語

処理の新しい応用」シンポジウム, pp.141-142.
9. 白井諭, 池原悟, 河岡司 (1993.5.20-21). 日英機械翻訳における原文自動書き替え型翻訳方式とその効果. 電子情報通信学会技術研究報告, NLC93-12, pp.9-16.
10. 白井諭, 横尾昭男, 池原悟, 木村淳子, 小見佳恵 (1994.7.21-22). 日本語従属節の依存構造に着目した係り受け解析. 情報処理学会研究報告, 94-NL-102-9,pp.65-72.
11. 池原悟, 白井諭, 横尾昭男, Francis Bond, 小見佳恵 (1994.7.21-22). 日英機械翻訳における利用者登録語の意味属性の自動推定. 情報処理学会研究報告, 94-NL-102-10,pp.73-80.
12. 小倉健太郎, 白井諭, 池原悟 (1995.3.7). 日英機械翻訳の副詞翻訳. 電子情報通信学会技術研究報告, NLC94-44, pp.1-8.
13. 白井諭, 池原悟, 横尾昭男, 木村淳子 (1995.5.12). 階層的認識構造に着目した日本語従属節間の係り受け解析の方法とその精度. 電子情報通信学会技術研究報告, NLC95-1, pp.1-8.
14. 池原悟, 白井諭, 河岡司 (1995.5.12). 大規模日本語コーパスからの連鎖型および離散型の共起表現の自動抽出法. 電子情報通信学会技術研究報告, NLC95-3, pp.17-24.
15. 白井諭, 池原悟, 横尾昭男, 井上浩子 (1995.11.17-18). 日英機械翻訳に必要な結合価パターン対の数とその収集方法. 情報処理学会研究報告, 95-NL-110-7, pp.43-50.
16. 高橋大和, 白井諭, 藤波進, 池原悟, 上田洋美, 松島英之 (1996.7.18-19). 日英新聞記事の自動記事対応付け. 電子情報通信学会技術研究報告, NLC96-17, pp.55-62.
17. 内野一, 白井諭, 池原悟, 新田見緑 (1996.7.18-19). 置換えを用いた **n-gram** による言語表現の抽出. 電子情報通信学会技術研究報告, NLC96-18, pp.63-68.
18. 白井諭, 上田洋美, 兵藤富子, 横尾昭男, 池原悟 (1996.10.11). 日英機械翻訳のための結合価パターン対の作成支援処理. 電子情報通信学会技術研究報告, NLC96-34, pp.25-30.
19. 松尾義博, 白井諭 (1996.11.18-19). 発音情報を用いた訳語対の自動抽出. 情報処理学会研究報告, 96-NL-116-15, pp.101-106.
20. 宮崎正弘, 池原悟, 横尾昭男, 白井諭 (1997.7.24-25). 日英機械翻訳のための意味属性体系. 電子情報通信学会技術研究報告, NLC97-12, pp.29-36.
21. 横尾昭男, 宮崎正弘, 池原悟, 白井諭, 阿部さつき (1997.7.24-25). 日英機械翻訳のための単語辞書. 電子情報通信学会技術研究報告, NLC97-13, pp.37-44.
22. 白井諭, 横尾昭男, 中岩浩巳, 池原悟, 宮崎正弘 (1997.7.24-25). 日英機械翻訳のための構文辞書. 電子情報通信学会技術研究報告, NLC97-14, pp.45-52.
23. 白井諭, 池原悟, 相澤弘, 鳴海武史, 横尾昭男 (1997.11.20-21). 結合価パターン対作成のための日英対訳用例文の収集. 情報処理学会研究報告, 97-NL-122-1,pp.1-6.
24. 白井諭, 大山芳史, 渡邊いづみ, 赤迫佐和子, 高橋直美, 石崎俊 (1997.11.20-21). 英単語に対する述語性の連語的日本語訳語の分析. 情報処理学会研究報告, 97-NL-122-2,pp.7-12.
25. 春野雅彦, 白井諭, 大山芳史 (1997.11.25-27). 決定木を用いた日本語係り受け解析. 自然言

語処理シンポジウム「実用的な自然言語処理に向けて」,
http://www.csl.sony.co.jp/person/nagao/nlsym97/index.html.
26. 中井慎司, 池原悟, 白井諭 (1998.5.15).「の」型名詞句における係り受け規則の自動生成法. 電子情報通信学会技術研究報告, NLC98-3, pp.45-51.
27. 中井慎司, 池原悟, 白井諭 (1998.11.5-6).「の」型名詞句における品詞情報と意味情報を併用した係り受け規則の自動生成. 情報処理学会研究報告, 98-NL-128-7,pp.45-51.
28. 白井諭, 大山芳史, 池原悟, 宮崎正弘, 横尾昭男 (1998.11.26-27). 日本語語彙大系について. 情報処理学会研究報告, 98-IM-34-9, pp.47-52.
29. 高橋大和, 松尾義博, 畑山満美子, 古瀬蔵, 白井諭 (1999.2.1-3). 新聞記事を対象とする日英対訳コーパスの作成状況.「言語資源の共有と再利用」シンポジウム,
http://www.carc.aist.go.jp/nlwww/sympo99/takahashi/index.html.
30. 白井諭 (1999.2.1-3). 結合価パターン対の網羅的収集に向けて －日英機械翻訳の観点からー.「言語資源の共有と再利用」シンポジウム,
http://www.kecl.ntt.co.jp/icl/mtg/members/shirai/nlpsym99.html.
31. 白井諭 (1999.3.19-20). ことばの組み合わせっていくつあるの**?** －日英機械翻訳のための単文の結合価パターン対の収集ー. 人工知能学会, SIG-LSE-9901-(8), pp.59-66.
32. 畑山満美子, 松尾義博, 白井諭 (2001.1.25-26). 重要語句抽出による新聞記事自動要約. 情報処理学会研究報告, 01-NL-141-16, pp.95-101.
33. 竹沢寿幸, 白井諭, 大山芳史 (2001.1.25-26). バイリンガル旅行会話コーパスに見られる話し言葉の特徴分析. 情報処理学会研究報告, 01-NL-141-22, pp.137-144.
34. 白井諭, 山本和英, Kyonghee Paik (2001.3.15-16). 対訳辞書作成のための英訳辞書の照合. 電子情報通信学会技術研究報告, TL2000-36/NLC2000-71, pp.17-24.
35. 白井諭, 山本和英 (2001.3.30). 換言事例の収集 －機械翻訳における多様性確保の観点からー. 言語処理学会第 7 回年次大会 ワークショップ論文集, 2L, pp.3-8.
36. 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 (2002.12.6). 等価的類推思考の原理による機械翻訳方式. 電子情報通信学会技術研究報告, TL2002-34, pp.7-12.
37. 池原悟, 佐良木昌, 宮崎正弘, 池田尚志, 新田義彦, 白井諭, 村上仁一, 徳久雅人 (2003.3.6-7). 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, TL2002-48, pp.1-6.
38. 衛藤純司, 池原悟, 池田尚志, 新田義彦, 柴田勝征, 宮崎正弘, 白井諭 (2003.5.26-27). 意味類型構築のための文接続表現の体系化について. 情報処理学会研究報告, 03-NL-155-6, pp.31-38.
39. 白京姫, 中岩浩巳, 白井諭 (2003.8.29). 言語的類似性を最大利用した直接翻訳方式. 電子情報通信学会技術研究報告, NLC2003-21, pp.37-42.

[特許]

1. 日本語文節間係り受け解析装置. 第 2021736 号（登録 1995.6.7, 白井諭）.
2. 日本語文節間係り受け解析装置. 第 2021737 号（登録 1995.6.7, 白井諭）.
3. 日本語文節間係り受け解析装置. 第 2035515 号（登録 1995.7.31, 白井諭）.
4. 日本語文節間係り受け解析装置. 第 2504447 号（登録 1996.4.2, 白井諭）.
5. 日本語文節間係り受け解析装置. 第 2504449 号（登録 1996.4.2, 白井諭）.
6. 自然言語解析システム. 第 2770555 号（登録 1998.4.17, 白井諭）.
7. 自然言語自動翻訳装置. 第 2804947 号（登録 1998.7.24, Francis Bond, 白井諭, 横尾昭男, 小倉健太郎）.
8. 自然言語自動翻訳装置. 第 2877608 号（登録 1999.1.22, 白井諭, Francis Bond, 横尾昭男, 内野一）.
9. 自然言語自動翻訳方式. 第 2915113 号（登録 1999.4.16, 中岩浩巳, 白井諭, 横尾昭男, Francis Bond）.
10. 自然言語翻訳装置. 第 2935928 号（登録 1999.6.4, 白井諭, 横尾昭男）.
11. 電子化辞書検索方法. 第 3025847 号（登録 2000.1.28, 内野一, 坂間保雄, 白井諭）.
12. 自然言語自動翻訳方法. 第 3206816 号（登録 2001.7.6, 内野一, 横尾昭男, 白井諭）.
13. 訳語対抽出装置. 第 3282789 号（登録 2002.3.1, 松尾義博, 白井諭）.
14. 自然言語自動翻訳装置. 第 3287068 号（登録 2002.3.15, 松尾義博, 白井諭, 横尾昭男）.
15. 自然言語翻訳装置. 第 3345763 号（登録 2002.9.5, 松尾義博, 白井諭, 横尾昭男）.
16. 利用者辞書作成支援装置. 第 3393494 号（登録 2003.1.31, 白井諭, 横尾昭男, Francis Bond, 池原悟）.

[著書]

1. Yoshihiro Matsuo, Satoshi Shirai, Akio Yokoo & Satoru Ikehara (1997.6). **Direct parse tree translation in cooperation with the transfer method**. *New Methods in Language Processing (Studies in Computational Linguistics)*, Daniel Jones & Harold Somers (eds.), Chapter 18, pp.229-238, UCL PRESS, ISBN 1-85728-711-8.
2. 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（編）(1997.9.26). 日本語語彙大系. 岩波書店, ISBN 4-00-009884-5.
3. 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（編）(1999.9.24). 日本語語彙大系 **CD-ROM** 版. 岩波書店, ISBN 4-00-130101-6.
4. Francis Bond & Satoshi Shirai (2003.6). **A hybrid rule and example-based method for**

machine translation. *Recent Advances in Example-Based machine translation*, Michael Carl & Andy Way (eds.), Chapter 7, pp.211-224, TEXT, SPEECH AND LANGUAGE TECHNOLOGY (Volume 21), Kluwer Academic Publishers, ISBN 1-4020-1400-7 or 1-4020-1401-5.