頑健な英日機械翻訳システム実現のための原文自動前編集

吉見　毅彦[†]　佐田いち子[†]　福持　陽士[†]

本稿では，従来の機械翻訳システムの構文解析能力を越える倒置や挿入などを含む文に対して頑健な処理を実現するための一手法として，形態素解析と簡単な構文解析によって得られる情報に基づいて原文を書き換える自動前編集手法を示す．原文書き換え系を既存システムに追加することによって，1) より品質の高い翻訳がシステムの既存部分にほとんど変更を加えることなく得られるようになるだけでなく，2) 構文解析の負担が減少するためシステム全体としての効率化が実現できる．実際，提案手法を我々の英日機械翻訳システム Power E/J に組み込み，新聞記事を対象として実験を行なったところ，1) 書き換え規則が適用された 330 文の 78.8%にあたる 260 文の翻訳品質が改善され，2) 書き換えを行なった場合の翻訳速度は行なわない場合の速度の 1.12 倍になった．

キーワード:　原文書き換え，自動前編集，不適格文，頑健性，機械翻訳

Automatic Preediting of English Sentences for a Robust English-to-Japanese MT System

Takehiko Yoshimi[†] , Ichiko Sata[†] and Yoji Fukumochi[†]

As a means of allowing for robust processing of such linguistic phenomena as inversion, ellipsis, parenthesis and emphasis, which are liable to prevent a syntactic parser from generating appropriate syntactic structures, this paper shows a method of automatically preediting sentences, based on information obtained by morpholexical and simple syntactic analysis. Addition of a preediting module to the existing system makes it possible 1) to generate better translations, which would not otherwise be generated, with little or no changes to the existing parts of the system, and 2) to reduce the load of syntactic analysis, thus enhancing the efficiency of the whole system. We have incorporated the proposed method into our English-to-Japanese machine translation system Power E/J, and carried out an experiment with sentences in news wire articles. The incorporation of the preediting module has satisfactorily 1) improved the quality of translations for the 260 sentences out of rewritten 330 ones (78.8%), and 2) marked up the speed of 1.12 times as fast as the system without the module.

KeyWords: *Rewriting, Automatic Preediting, Ill-formedness, Robustness, Machine Translation*

† シャープ (株) 情報システム事業本部, Information Systems Group, SHARP Corporation

1　はじめに

様々な状況で利用される機械翻訳システムが直面する現実の文には，システムが持つ言語知識では適切に解析できない様々な言語現象が現れる．このような現象を含む文は，人間にとっても適格でない (が理解できる) 絶対的不適格文と，人間にとっては適格であるがシステムの処理能力を越えている相対的不適格文に分けられるが，両者を適切に扱える頑健なシステムが求められている (松本裕治 今一修 1994)．絶対的不適格現象のうち語句の欠落や主語述語の不一致などの構文レベルの現象へ対処することを目的とした手法としては，部分解析法 (Imaichi and Matsumoto 1995) や制約緩和法 (Mellish 1989; 加藤恒昭 1995) などがこれまでに提案されている．

他方，我々は，相対的不適格文への対処に焦点を当て，機械翻訳システムの翻訳品質の向上を目指している．以降本稿では紛れない限り，相対的不適格文を単に不適格文と呼ぶ．構文レベルの不適格文すなわちシステムの解析能力を越えた構文構造を持つ文を扱うための代表的な手法には，1) 対象テキストの分野を限定した専用文法を用いる手法 (相沢輝昭, 加藤直人, 鎌田雅子 1996) や，2) 原文を書き換える手法 (金淵培 江原暉将 1994; 成田一 1994; 佐川雄二, 大西昇, 杉江昇 1994; 白井諭, 池原悟, 河岡司, 中村行宏 1995; 加藤輝政, 小川清, 佐良木昌 1997) などがある．また，後者の手法に関連して，原文とそれを人間が書き換えた結果とを比較した差分から原文書き換え規則を学習する手法 (山口昌也, 乾信雄, 小谷善行, 西村恕彦 1998) も示されている．(1) と (2) の手法の設計方針は，システムの既存部分の変更を避け，新たな処理系を追加するという点で共通しているが，以下の点で異なっている．前者の手法では，システムの既存部分による処理は，可能な場合には，新たに追加した処理系による処理によって代行される．すなわち，新たな処理系による解析 (分野依存の専用文法による解析) が成功した場合には，既存の処理系による解析 (汎用文法による解析) は実行されない．これに対して後者では，新たに追加した処理系は既存部分の前処理系と位置付けられる．

原文書き換えによる手法は，書き換えを構文解析の前に行なうか後に行なうかによって二つに分けられる．構文解析後に行なう場合 (佐川雄二他 1994; 白井諭他 1995) [1]は，構文情報が得られているため，構文解析前すなわち形態素解析後に行なう場合に比べてより翻訳品質の高いシステムが実現できる可能性がある．しかし，実用的な機械翻訳システムにおいて原文書き換えの実行を構文解析終了後まで遅らせることは，処理効率の点では望ましくない．なぜならば，入力文全体を覆う構文構造が生成できず構文解析に失敗すること[2]が判明するのは構文解析規則をすべて適用し終えた後であるが，実用的な機械翻訳システムでは構文解析規則の規模は非常に大きくなっており，構文解析に要する時間は解析全体に要する時間の大半を占めているため，

1 白井らは，文献 (Shirai, Ikehara, Yokoo, and Ooyama 1998) で，一部の書き換えを構文解析前に行なうように拡張を施しているが，書き換え規則の多くは構文解析後に適用される．
2 以降本稿では，入力文全体を覆う構文構造が生成できないことを構文解析の失敗と呼ぶ．

構文解析後の書き換えは処理の効率化につながりにくいからである．これに対して，構文解析が失敗しないようにあらかじめ原文を書き換えれば，すべての構文解析規則の適用が試みられる可能性は低くなるため，システム全体として効率の良い処理が実現できる．また，構文情報が (ほとんど) 得られていない時点で行なう書き換えがどの程度有効であるかを明らかにすることも重要である．

このような観点から本稿では，形態素解析で得られる情報と通常よりも簡単な構文解析[3]で得られる情報に基づいて原文書き換えを構文解析前に行なうことによって翻訳品質と共に翻訳速度を改善する手法を示す．以下，本稿で扱う書き換え対象を 2 節で整理する．次に 3 節で原文書き換え系の処理枠組について説明する．4 節では，原文書き換え系を既存の英日機械翻訳システムに組み込み，システムの性能向上にどの程度貢献できるかを実験によって検証する．5 節では関連研究との比較を行なう．

2　書き換え対象文

我々の従来システムにとっての不適格現象には様々なものがあるがそれらをすべて一度に扱うことは容易ではない．このため本研究では，出現頻度が高い現象や翻訳品質の改善度が大きい現象を含む文を書き換え対象として優先的に選ぶことにする．書き換え対象を選定するために，英文法書の例文や新聞記事 (Lewis 1997) を我々のシステムで処理し，構文解析に失敗した文のうち 558 文についてその原因を分析した．構文解析に失敗する現象は 595 箇所で見られた．このうち 238 箇所は，綴り誤りなどの絶対的不適格現象や，辞書や形態素解析系の不備によるものであった．残り 357 箇所のうち 108 箇所が省略現象によるもの，61 箇所が倒置によるもの，41 箇所が挿入語句によるものであった．また，特殊構文を含まないが文が長いために構文解析に失敗する文が 26 文存在した．

この分析結果に基づいて，表 1 に示すパターンへ対処することを目的とした[4]．表 1 のパターンで，構文解析に失敗した 558 文に現れたすべての倒置，省略，挿入を網羅しているわけではない．パターン 7 ないし 10 は，複雑で長い部分を動詞の後方に置くための前置詞句の前置であるが，ここでは倒置とみなす．表 1 において，上付き記号?は任意項を，{ } は選択項をそれぞれ表し，斜字体の語句は省略されている語句を意味する．

表 1 のパターンは我々の従来システムにとっての不適格現象であり，他のシステムの中には適切に処理できるものも存在すると考えられる．しかし，例えば日本電子工業振興協会の自然言語処理技術委員会で編集された翻訳品質評価用テストセット (日本電子工業振興協会 1995) で取り扱いが重要な特殊構文として取り上げられているように，これらを適切に処理することは

3 具体的には，3.2.1 節で述べる手続きによる処理を指す．

4 表 1 は対象とするパターンの概略であり，実際の書き換え規則の適用条件部にはより厳密な条件を記述している．また，強調構文は構文解析に失敗するわけではないが重要な構文であるので，書き換え対象に含めた．

表 1　書き換え対象パターン

分類	#	パターン
倒置	1.	[副詞句? { 分詞 | 形容詞 } 前置詞句? 助動詞? be 動詞]
	2.	[副詞句? 前置詞句 助動詞? be 動詞]
	3.	[{nor|neither} 助動詞 名詞句 原形動詞]
	4.	[{nor|neither} have 名詞句 過去分詞]
	5.	[{nor|neither} be 動詞 名詞句]
	6.	[should⟨ 文頭 ⟩]
	7.	[関係代名詞 前置詞句 動詞]
	8.	[名詞句 前置詞句 ⟨ 意味標識:TIME⟩ 動詞]
	9.	[助動詞 前置詞句 原形動詞]
	10.	[have 前置詞句 to 原形動詞]
省略	11.	[so 形容詞 *that* 代名詞]
	12.	[say *that* ··· and that ···]
	13.	[say *that* there 動詞]
	14.	[{double|twice} that *figure*]
挿入	15.	[, {all|most|much|···} of {it|them} ··· {,|.}]
	16.	[, but not 動詞 ··· ,]
	17.	[名詞句 and 名詞句 , 名詞句 , 動詞]
	18.	[名詞句 , {which|who} ··· , 動詞]
強調	19.	[it be 動詞 not? 名詞句 {which|who}]

一般の機械翻訳システムにとっても重要な課題である．また，これらの現象のうち，倒置，省略，挿入は，英日機械翻訳システムの一般利用者が日々接することが多いテキストの一つである英字新聞記事に比較的頻繁に現れる表現である (上野田守 布施敏夫 1978; 堀内克明 1979; 富田春生 1994) ため，これらを適切に処理する必要性は高い．

本節では，これらの現象 (江川泰一郎 1964; 安井稔 1982; Greeenbaum and Quirk 1990) を含む文をどのように書き換えれば翻訳品質が改善されるかを検討する．以降，従来システムとは我々の従来システムを指す．

2.1　倒置

倒置のうち分詞を中心とする叙述部の倒置や否定表現が冒頭に置かれた文における倒置などを扱う．例えば文 (E1) は過去分詞が文頭に現れているために，構文解析に失敗し (J1) のような出力しか得られない[5]．しかし，文 (E1) の先頭に “what is” という語句を追加して文 (E1') のように書き換えると，従来システムにとって適格文となり文 (J1') のような翻訳が得られる．

(E1)　Affiliated is the parent company of Globe Newspaper Co.

(J1)　◇加入した ‖ Globe Newspaper Co の親会社である。

(E1')　What is affiliated is the parent company of Globe Newspaper Co.

5 記号◇で構文解析が失敗したことを表し，記号 ‖ で区切られた区間が部分的な構造にまとめられたことを表す．

(J1')　合併されるものは、Globe Newspaper Co の親会社である。

文 (E2) に見られるように，否定の副詞 “neither” が冒頭に置かれた節では主語と助動詞の倒置が生じる．このような節に対しては，“neither” を “and also” に書き換えて，助動詞 “have” を否定形にし主語の後方に移動する．このような処理の実現には，助動詞の移動先を決定しなければならないため，主語になる名詞句この場合 “ the government regulators” を認識する必要がある．このため，書き換え規則の適用条件部には簡単な構文構造を認識するための手続きも記述する．この手続きに関しては 3.2.1 節で述べる．

(E2)　Neither have the government regulators indicated that there will be a problem.

(J2)　◇どちらも、政府取締官を示された状態にしない ‖ そこのそれは、問題であろう

(E2')　And also the government regulators have not indicated that there will be a problem.

(J2')　そしてまた、政府取締官は、問題があるであろうことを示さなかった。

2.2　省略

接続詞の省略など構文的知識に依存する省略のいくつかを扱う．文 (E3) は，二つの被伝達節が “and” で連結されているが，一つ目の被伝達節を導く “that” が明示されていない．これが原因で構文解析に失敗するが，“said” の直後に “that” を補えば構文解析の失敗は避けられるようになる．

(E3)　Sprinkel said the fall of the dollar had substantially restored U.S. cost competitiveness and that the deterioration of the U.S. trade balance appeared to have abated.

(J3)　◇Sprinkel によれば、ドルの低下は、米国のコスト競争、及び、それを大幅に回復した ‖ 米国の貿易収支の悪化は、減少したために、現れた。

(E3')　Sprinkel said that the fall of the dollar had substantially restored U.S. cost competitiveness and that the deterioration of the U.S. trade balance appeared to have abated.

(J3')　Sprinkel は、ドルの低下が米国のコスト競争を大幅に回復したということ、そして、米国の貿易収支の悪化が減少したように思われるということを言った。

また，文 (E4) で “so” と相関関係にある “that” が省略されていることは，“so” の直後の形容詞に代名詞が後続していることを手がかりにすれば検出できる．

(E4)　The 2.5 pct discount rate is so low it is politically impossible to cut it further.

(J4)　2.5 パーセント割引率は、そのようにそれが政治上まで不可能である安値が更にそれを切ったことである。

(E4') The 2.5 pct discount rate is so low that it is politically impossible to cut it further.

(J4') 2.5 パーセント割引率は、更にそれを切ることが政治上不可能であるほど低い。

2.3 挿入

挿入句は文中の他の語句と特に構文的な関係を持つことなく現れ，コンマやダッシュや括弧などで区切られる．括弧で囲まれた挿入句を発見することは比較的容易であるが，コンマで囲まれている場合にはコンマの他の用法との区別を行なう必要があり，挿入句の正確な認識は容易ではない (武田紀子 1995)．しかし，コンマだけでなくその前後の語句も手がかりとし，表 1 のようなパターンとして捉えれば，挿入句の発見はより正確に行なえる．例えば文 (E5) において，一つ目のコンマの直後に存在する “some of them” のような特徴的な語句や二つ目のコンマの直後に存在する (助) 動詞に着目すればよい．このような手がかりに基づいて認識した挿入句を括弧で囲めば，既存の構文解析系による処理が成功するようになる．ただし，文 (E5) は挿入だけでなく be 動詞の省略も含んでいるため，“them” の直後に “are” を補う必要がある．

(E5) Their separate proposals, some of them conflicting, will be woven by House Democratic leaders into a final trade bill for a vote by the full House in late April.

(J5) ◇それらの個別の提案、‖ いくらかの ‖ それら ‖ 対立する、‖ 下院の民主党のリーダーによって 4 月下旬の下院本会議による投票のための最終の貿易手形に織られるであろう。

(E5') Their separate proposals (some of them are conflicting) will be woven by House Democratic leaders into a final trade bill for a vote by the full House in late April.

(J5') それらの個別の提案 (それらのうちのいくらかが対立している) は、下院の民主党のリーダーによって 4 月下旬の下院本会議による投票のための最終の貿易手形に組み立てられるであろう。

2.4 強調

強調のための言語的手段には，2.1 節で述べた倒置文の他に感嘆文や修辞疑問文や分裂文などがある．このうち従来システムでは文 (E6) のような分裂文を適切に処理することができないが，これを文 (E6') のように書き換えれば翻訳が改善される．文 (E6') は，文 (E6) に対して，“it” を削除し，“which” を “what” に置換し，焦点の名詞句とその直前の be 動詞を “which” 節の後方へ移動するという三操作を行なうことによって得られる．

(E6) However, he added that in the end, it was market forces which prevailed.

(J6) しかしながら、彼は、結局それが普及していた市場諸力であるとつけ加えた。

(E6')　However, he added that in the end, what prevailed was market forces.

(J6')　しかしながら、彼は、結局普及していたものが市場諸力であるとつけ加えた。

3　原文書き換え系

3.1　原文書き換えの枠組

本節で述べる原文書き換え系を組み込んだ機械翻訳システムにおける解析の流れを図 1 に示す．このシステムでは，形態素解析終了後に書き換えを実行した後，書き換えた部分の形態素解析を行ない，入力文全体の形態素解析結果を構文解析系に送る．一度目の書き換え結果に対して全体を覆う構文構造が生成できず構文解析に失敗した場合，処理の制御は原文書き換え系に戻る．再度書き換えを行なう場合には，各書き換え規則に記述されている規則の信頼度 (後述) に従って，一度目の書き換えでは用いなかった規則を新たに適用したり，逆に一度目の処理で行なった書き換えを取り消したりする[6]．

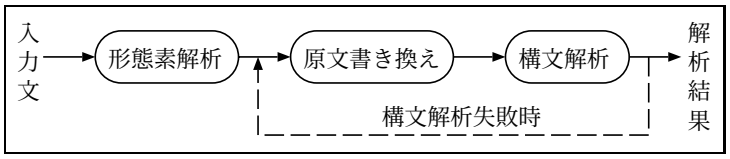

図 1　解析の流れ

原文書き換え系での処理は，形態素解析結果に対して先頭から順に書き換え規則の適用条件との照合を行なっていき，適用条件が満たされる部分を順次書き換えていく．

3.2　書き換え規則の形式

書き換え規則には，次に示すように，適用条件と書き換え操作の他，制御情報として適用抑制規則集合と信頼度を記述することができる．

(識別番号，適用条件，書き換え操作，適用抑制規則集合，信頼度)

3.2.1　適用条件部

適用条件は，ある指定された属性を持つ語句がある位置あるいは区間に存在するかどうかを調べる手続きの論理積，論理和の形で表現される．入力文中での書き換え対象候補の出現位置は次のように表す．まず着目語の出現位置を変数 p で表し，$p+n$ で着目語の n 語後方 (文末側)

6 二度目の構文解析に失敗した場合には，断片的な構文構造を解析結果とする．

の語を，$p-n$ で着目語の n 語前方 (文頭側) の語を指す．コンマや終止符なども語とみなす．特別な記号として文末の語を指すドル記号\$を用いる．この他，コンマが書き換え対象を発見するための重要な手がかりとなることが多い (Jones 1994) ため，com_fwd_m で着目語の後方に存在する m 番目のコンマを指し，com_bwd_m で前方の m 番目のコンマを指すことにする．例えば word_class(com_fwd_1 + 2, noun) という手続きは，着目語に最も近い後方のコンマの二語後方に，語種 (品詞) 候補として名詞を持つ語が存在するとき真を返し，存在しないとき偽を返す．

適用条件部に記述する手続きは主に語の形態素語彙属性を調べるものであるが，特別な手続きとして，入力文のある部分を節や名詞句と解釈できるかという構文的な属性を調べる手続きを記述する．ただし，簡単な構文解析しか行なわない方針であることから，節の認識は次のような手順で行なう．

> 処理対象の先頭から順に述語になり得る定形動詞を探していく．もし見つかれば，その述語候補と人称・数が一致する名詞を主辞とする名詞句がその前方に存在するかどうかを調べる．もしそのような名詞句が存在すれば，それを主語とみなし，節が存在するものとする．

ここで，名詞句を認識する手続きは次のような簡単な構造を検出するものである．上付き記号?と+はそれぞれ一回以下，一回以上の出現を意味する．

$$\begin{aligned} \text{名詞句} &= \text{名詞句}\,0\ (\text{前置詞 名詞句}\,0)^{?} \\ \text{名詞句}\,0 &= (\text{副詞}^{?}\ \{\ \text{形容詞} \mid \text{過去分詞} \mid \text{現在分詞}\ \})^{?}\ \text{名詞}^{+} \end{aligned}$$

3.2.2　操作部

書き換え操作には，語句の追加・削除・置換・移動，文の分割，前編集記号の付加がある．前編集記号の付加は，語の形態素語彙属性を指定したり，節や句の範囲や従属先を指定したりするためのものである．語句の語彙属性や従属先の指定によって解釈の曖昧性が減るため，解析の品質と速度の向上が期待できる．

3.2.3　適用抑制規則集合

ある規則 R に与えられている適用抑制規則集合は R の適用を抑える他の規則に関するメタ条件を表し，規則 R はその適用抑制規則集合に記述されている識別番号の規則が既に適用されている場合には適用されない．規則 R の適用抑制規則集合には，R の書き換え対象と重複する部分を書き換えようとする規則だけでなく，書き換え対象が R のものと重複しない規則を含めてもよい．

3.2.4　信頼度

規則には，その信頼性が高く規則の適用によって翻訳品質が向上することがほぼ確実な規則もあれば，信頼性があまり高くない規則もある．信頼度は，このようなことを考慮して，信頼性があまり高くない規則による悪影響を抑えるために設定したものである．各規則には，その信頼性に応じて A，B，C のいずれかの信頼度を与える．信頼度 A の規則は最初の構文解析の前に適用し，構文解析に失敗してもこの規則による書き換えは取り消さない．規則に信頼度 A を与えるのは，この規則を適用しないと構文解析に失敗することがほぼ確実であり，たとえこの規則によって書き換えた表現の構文解析に失敗して断片的な構文構造しか得られなかったとしても，この規則を適用しない場合の (断片的な) 構文構造から生成される翻訳よりも高い品質の翻訳が生成されると期待される場合である．信頼度 B の規則は最初の構文解析の前に適用するが，最初の構文解析に失敗した場合，この規則による書き換えは取り消す．信頼度 C の規則は最初の構文解析の前には適用せず，最初の構文解析に失敗した場合に初めて適用する．各規則にどの信頼度を与えるかは実験を繰り返して経験的に決定する．

3.3　書き換え規則の例

書き換え規則の例として，挿入と省略を含む文 (E5) に対処するための規則と分裂文 (E6) に対処するための規則を図 2 に示す．図 2 の最初の規則が挿入と省略のための規則であり，二つ目が分裂文のための規則である．ただし，これらは，理解を容易にするため実際の規則を簡単化したものである．

最初の規則の適用条件が満たされるのは，現在着目している位置に “some of them” が存在し，着目語の直前がコンマであり，着目語に最も近い後方のコンマの直後に動詞が存在し，二つのコンマに挟まれた区間に “some of them” を主語とする述語が存在しないときである．このとき，コンマを括弧に書き換え，“some of them” の直後に be 動詞 “are” を挿入する．

二つ目の規則は，現在着目している位置に代名詞 “it” が存在し，着目語の直後が be 動詞の肯定形あるいは否定形であり，その直後に名詞句が存在し，その名詞句の直後に “which” が存在するときに適用する．

4　実験と考察

原文書き換えの有効性を検証するために，提案手法を我々の従来システムに組み込み，翻訳品質がどの程度向上するかを調べる実験と翻訳速度がどの程度向上するかを調べる実験を行なった．

```
(/* 識別番号 */
 36,
 /* 適用条件 */
 (word(p, p+2, "some of them") == TRUE &&
  word(p-1, ",") == TRUE &&
  word_class(com_fwd_1+1, verb) == TRUE &&
  subject_predicate(com_bwd_1+1, com_fwd_1-1) == FALSE),
 /* 書き換え操作 */
 (substitute(com_bwd_1, "("),
  substitute(com_fwd_1, ")"),
  insert(p+2, "are")),
 /* 適用抑制規則集合 */
 (),
 /* 信頼度 */
 A
)

(/* 識別番号 */
 42,
 /* 適用条件 */
 (word(p, "it") == TRUE &&
  word_class(p+1, be) == TRUE &&
  (noun_phrase(p+2, q) == TRUE ||
   word(p+2, "not") == TRUE && noun_phrase(p+3, q) == TRUE) &&
  word(q+1, "which") == TRUE),
 /* 書き換え操作 */
 (remove(p),
  substitute(q+1, "what"),
  move(p+1, q, $-1)),
 /* 適用抑制規則集合 */
 (),
 /* 信頼度 */
 B
)
```

図 2　書き換え規則の例

4.1　実装

今回の実験のために実装した書き換え規則の総数は 35 規則であり，その内訳は倒置，省略，挿入，強調用がそれぞれ 18，6，8，3 規則である[7]．規則の適用条件部には，形態素語彙属性を調べる手続きと構文的な属性を調べる手続きの二種類が記述されている．構文的な属性を調べる手続きは，3.2.1節で述べた，簡単な名詞句を認識する手続き noun_phrase() と，これを利用して節を認識する手続き subject_predicate() である．形態素属性を調べる手続きは，単語の語形，語種 (品詞)，意味標識などの照合を行なうものである．形態素属性を調べる手続きの一覧

7 規則作成に要した時間的，人的コストも提案手法の有効性を判断する上で重要なファクターであるとの指摘を査読者の方より受けたが，記録がなく記載できない．

を付録の表 6 に示す.

各規則の適用条件部に含まれる平均手続き数を表 2 に示す．形態素属性を調べる手続きは，表 2 によれば，実装した規則全体の平均で 10.1 個含まれている．最も多い規則では 16 個，最も少ない規則では 3 個である．構文属性を調べる手続きについては，平均で 0.9 個，最も多い規則で 3 個，最も少ない規則では 0 個である．構文属性を調べる手続きを全く含まない規則は 10 規則存在する.

表 2　適用条件部を構成する平均手続き数

	倒置	省略	挿入	強調	全体
形態素属性	10.6	10.0	10.8	6.0	10.1
構文属性	1.2	0.3	0.5	1.0	0.9

4.2　翻訳品質評価実験

4.2.1　実験方法

評価実験には新聞記事 (Lewis 1997) から抽出した 6182 文を用いた．これらのうち 1282 文は従来システムでの構文解析に失敗するものであり，残り 4900 文は失敗しないものである．評価文集に含まれる文の長さは，最も短いもので 4 語，最も長いもので 63 語，平均では 27.7 語であった．評価文集には書き換え規則作成時に参照した文も含まれているため，今回の実験は完全ブラインドテストではない[8].

規則記述の方針として，書き換えるべき表現が書き換えられない (再現率が上がらない) ことにはあまり注意せず，書き換えるべきでない表現を書き換えてしまう誤りの発生を極力抑える (適合率を上げる) ことにした．このため，評価では適合率のみに着目し，書き換えるべきでない表現が誤って書き換えられていないかを調べた．次に，規則が正しく適用されている文について，原文書き換え系を組み込んだ場合と組み込まない場合とで次の二点について比較した.

解析品質　原文書き換えを行なうことによって，構文解析失敗の頻度がどの程度減少するか．失敗とは，1 節で述べたように，入力文全体を覆う構文構造が生成できないことを意味する.

翻訳品質　翻訳品質がどの程度向上するか．評価値は，品質の向上・若干向上・低下・若干低下・同等のいずれかとした．評価の実施は第三者一名に依頼した.

8 本稿では，提案手法の技術的限界を知ることを主な目的としているため，ユーザの立場からの評価であるブラインドテストは今後の課題とする.

4.2.2　実験結果

実験に用いた 6182 文のうち書き換え規則が適用された文は 5.3%にあたる 330 文であった．書き換えるべきでない表現に規則が誤って適用された文は存在しなかった．二つ以上の規則が適用された文は存在しなかった．

書き換えられた 330 文の構文解析品質の改善度を表 3 に示す．表 3 によれば，330 文の 55.8%にあたる 184 文について，失敗していた構文解析が成功するようになっている．ここで，「成功」とは，入力文全体を覆う構文構造が生成できたことを意味しており，人間にとって正しい解釈が生成できたことを必ずしも意味しない．原文書き換えを行なっても依然として構文解析に失敗している 33 文の内訳は，今回書き換え対象としなかった等位構造などの相対的不適格現象を含むものが 10 文，記述した書き換え規則で捉えられなかった挿入や省略を含むものが 8 文，綴り誤りなどを含む絶対的不適格文が 7 文などであった．原文書き換えによって「成功」から「失敗」に悪化した文は存在しなかった．なお，表 3 は不適格現象ごとの集計ではなく書き換え規則ごとの集計である．例えば 2 節で挙げた文 (E5) は挿入と省略の二つの不適格現象を含むが，図 2 に示した一つの書き換え規則で書き換えられるため，挿入に関する規則としてのみ数えた．

表 3　構文解析品質の改善

	倒置	省略	挿入	強調	計
失敗 → 成功	93	58	33	0	184(55.8%)
失敗 → 失敗	3	17	13	0	33(10.0%)
成功 → 成功	23	86	1	3	113(34.2%)
成功 → 失敗	0	0	0	0	0(0.0%)
計	119	161	47	3	330(100%)

330 文についての翻訳品質改善度の評価結果を表 4 に示す．表 4 によれば，規則が適用された 330 文の 78.8%にあたる 260 文で品質改善が見られる．

表 4　翻訳品質の改善

	倒置	省略	挿入	強調	計
向上	62	64	25	3	154(46.7%)
若干向上	49	44	13	0	106(32.1%)
同等	5	41	8	0	54(16.4%)
若干低下	0	5	1	0	6(1.8%)
低下	3	7	0	0	10(3.0%)
計	119	161	47	3	330 (100%)

書き換え規則が正しく適用されているにも拘らず評価値が「低下」となった 10 文についてその原因を分析した[9]．省略に関する 7 文はすべて接続詞 “that” の省略を含むものであった．“that” の補完が文 (E3) のように品質改善につながることも多いが、文 (E7) のようにつながらないこともある．文 (E7) を翻訳した文 (J7) では，“that” は接続詞と解釈されずに形容詞と誤解釈されている[10]が，“and” 以降の節 “that ... tomorrow.” も被伝達節として正しく認識されている．これに対して文 (J7') では，被伝達節として認識されているのは二つの節のうち一つ目の節 “telephone ... today” だけである．文 (J7') は，“1500 hrs EST” などの時間表現が原語のまま残っているが，被伝達節の範囲が誤りであることは原文と照らし合わせなければ判明しないため誤解を招く危険な翻訳である．

(E7) The office said telephone confirmation of allotments must be received by 1500 hrs EST today and that secondary trading will begin at 0930 hrs EST tomorrow.

(J7) そのオフィスは言った。割当額の電話確認は、1500 hrs EST によって今日受け取られなければならず、そして、その第 2 の取引は、0930 hrs EST から明日始まるであろうと。

(E7') The office said that telephone confirmation of allotments must be received by 1500 hrs EST today and that secondary trading will begin at 0930 hrs EST tomorrow.

(J7') そのオフィスは、割当額の電話確認が 1500 hrs EST によって今日受け取られなければならないと言い、そして、その第 2 の取引は、0930 hrs EST から明日始まるであろう。

評価値が「低下」となった残りの 3 文は，文 (E8) のように，二つの節 “should ... this quarter” と “this action ... negotiations” の節境界を示すコンマが存在しないために，書き換え後の文 (E8') において節境界の認識を誤ったものである．文 (E8') を翻訳した文 (J8') では，“If ... urgency” が名詞節と解釈されているが，“if” で導かれる節が名詞節ではなく副詞節であることは，文頭の “should” を “if” に置き換える規則を文 (E8) に適用する段階で判明しているので，“if” の語種 (品詞) を副詞節接続詞と指定する前編集記号 “ca_” を付加する操作を書き換え規則に追加し，文 (E8") のように書き換えることは可能である．この修正によって文 (J8") のような翻訳が得られる．

(E8) Should Citicorp actually place the Brazilian loans in a non-performing category at the end of this quarter this action would serve to alleviate the urgency associated with the debt negotiations, he argues.

(J8) ◇シティコープ ‖ ブラジルのローンを終りの契約不履行のカテゴリに実際に置く ‖ の。今期にこの活動が負債交渉によって関連していた切迫を緩和するのに役立

9 原因は，実験に用いたシステムの既存部分の不備にあり，書き換え自体は文法的に正しい．
10 “that secondary trading” が名詞句と解釈され「その第 2 の取引」と翻訳されている．

つであろう、と彼は主張する。

(E8') If Citicorp actually place the Brazilian loans in a non-performing category at the end of this quarter this action would serve to alleviate the urgency associated with the debt negotiations, he argues.

(J8') シティコープがブラジルのローンを切迫を緩和するためにこの活動が役立つであろう今期の終りの契約不履行のカテゴリに実際に置くかどうかが負債交渉と結合した、と彼は主張する。

(E8") ca_If Citicorp actually place the Brazilian loans in a non-performing category at the end of this quarter this action would serve to alleviate the urgency associated with the debt negotiations, he argues.

(J8") シティコープがブラジルのローンを今期の終りの契約不履行のカテゴリに実際に置くならば、この活動が負債交渉と結合していた切迫を緩和するのに役立つであろう、と彼は主張する。

4.3 翻訳速度評価実験

4.3.1 実験方法

構文解析前に原文を書き換えれば，その分の処理の負担が増加する一方で，すべての構文解析規則の適用が試みられる可能性が低くなり構文解析の負担が減少する．このため，システム全体としては処理効率が向上すると予想される．この点を確認するために，原文書き換え系を従来システムに組み込んだ場合と組み込まない場合の翻訳速度を比較した．翻訳時間の測定は次の四種類の評価文集について行なった．

評価文集 1　品質評価実験に用いた 6182 文.

評価文集 2　評価文集 1 のうち構文解析に失敗する 1282 文.

評価文集 3　評価文集 1 のうち書き換え規則が適用された 330 文.

評価文集 4　評価文集 2 のうち書き換え規則が適用された 217 文.

4.3.2 実験結果

各評価文集について，原文書き換え系を組み込まない場合の一文当り平均の翻訳時間と，組み込んだ場合の一文当り平均の翻訳時間，さらに，前者の翻訳速度を 1 としたときの後者の翻訳速度を表 5 に示す．実験に用いた計算機の CPU は Pentium(R) II 400MHz，メモリは 128MB，OS は Windows(R) 98 である．翻訳システムは C 言語で記述されている．

評価文集 1 を対象とした実験の結果，原文書き換えを行なった場合の速度は行なわない場合の速度に対して 1.12 となっている．評価文集 1 において実際に書き換えられた文の数は入力文数の 5.3%に過ぎないが，このことを考慮すると翻訳速度向上への原文書き換えの貢献度は高い

と考えられる．

評価文集 3 を対象とした実験の結果より，書き換えるべき文がすべて書き換えられた場合には翻訳速度は 2.6 倍程度にまで向上するという一つの目安が得られた．また，構文解析が失敗することがあらかじめ判明している文だけを対象とした場合には，評価文集 1 や 3 を対象とした場合よりも大きな効果が現れることが確認された．

表 5　翻訳速度の比較

評価文集	原文書き換えなし (秒)	あり (秒)	速度比
1	1.20	1.07	1.12
2	2.63	2.01	1.31
3	3.93	1.46	2.69
4	5.27	1.51	3.49

原文書き換え系を組み込むことによってシステム全体の処理効率が向上したのは，具体的には次の二つの理由による．我々の構文解析系は二段階方式に基づいており，適格文用の構文解析規則を用いて解析を行なう機構と，この機構による通常の解析が失敗した時点で起動され，解析途中で生成された部分構造の中から発見的知識を用いて妥当なものを選び出すための別の機構を備えている．第一の理由は，システムにとっての不適格文が原文書き換え系によって適格文に書き換えられると，第二の機構による処理を実行する必要がなくなるからである．従って，構文解析前に原文書き換えを行なうことによる処理効率向上の効果は，我々のシステムに限らず，制約緩和法や部分解析法のように二段階方式で構文解析を行なっているシステムで一般に期待できる．

システムが想定していない言語現象を含む文の構文解析が失敗することは，第一の機構に記述されている規則をすべて適用し終えないと判明しない．これに対して，システムが想定している言語現象の文の解釈は，すべての規則を適用しなくても生成できる．速度向上のもう一つの理由は，原文書き換えによって，第一の機構で適用される規則の数すなわち生成される部分構造の数が減っていることである[11]．

今回実装した書き換え規則では実際には記述していないが，3.2.2 節で述べたように，本稿の原文書き換え系では書き換え操作として前編集記号を付加する操作を記述することができる．前編集記号の付加によって解釈の曖昧性が減るため，この操作を記述することによって解析速度がさらに向上する．例えば，文 (E8') では “if” に名詞節接続詞か副詞節接続詞かの曖昧性があるが，文 (E8") では副詞節接続詞に決定されるため，文 (E8') を文 (E8") に書き換えれば，さらに翻訳時間が短縮される．

11 適用される規則の数が具体的にどの程度減ったかの検証は本稿の範囲を越える．

5　関連研究との比較

原文書き換えを，形態素解析で得られる情報と通常よりも簡単な構文解析で得られる情報に基づいて行なう手法としては，金らの方法 (金淵培・江原暉将 1994) や加藤らの方法 (加藤輝政他 1997) がある．金らの方法は，長い日本語文の構文解析が失敗しやすいという問題に，長文を複数の短文に分割し，必要な場合には各短文に主語を補うことによって対処するものである．これに対して本稿の原文書き換え系では，文の分割だけでなく，語句の追加・削除・置換・移動，前編集記号の付加が可能であり，単に文を分割する場合よりも品質の高い翻訳が得られる．

加藤らは，英語の複文に着目してその編集方法を示しているが，書き換え方法の提案に留まっており評価結果は報告されていない．

6　おわりに

本稿では，一部の構文レベルの相対的不適格文を既存システムでも適切に扱えるように書き換えることによって頑健な処理を実現する手法を示した．この原文書き換え系を既存の英日機械翻訳システムの形態素解析系と構文解析系の間に組み込み，翻訳の品質と速度が改善されることを実験によって確認した．

倒置，省略，挿入，強調の現象にはそれぞれ様々なパターンがあるが，今回着目したパターンは比較的単純なものであり，記述した規則で多様なパターンを網羅しているわけではない．また，書き換えるべき表現が書き換えられないことにはあまり注意を払わなかった．今後，これらの点を考慮に入れた規則の拡張が必要である．

謝辞

英々変換系の初期の実装を行なって頂いたシャープ (株) 設計技術開発センターの関谷正明さんと，議論に参加頂いた英日機械翻訳グループの諸氏に感謝します．また，本稿の改善に非常に有益なコメントを頂いた査読者の方に感謝いたします．

参考文献

相沢輝昭, 加藤直人, 鎌田雅子 (1996). “外電経済ニュースの英日機械翻訳.” 情報処理学会論文誌, **37** (6), 1041–1048.

江川泰一郎 (1964). 英文法解説. 金子書房.

Greeenbaum, S. and Quirk, R. (1990). *A Student's Grammar of the English Language*. Longman. 池上嘉彦 他 訳. 現代英語文法 大学編, 紀伊國屋書店, 1995.

堀内克明 (1979). 時事英語. 朝日実務英語シリーズ. 朝日出版社.

Hornby, A. S. (1977). 英語の型と語法. オックスフォード大学出版局. 伊藤健三 訳注.

Imaichi, O. and Matsumoto, Y. (1995). "Integration of Syntactic, Semantic and Contextual Information in Processing Grammatically Ill-Formed Inputs." In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI95)*, pp. 1435–1440.

Jones, B. E. M. (1994). "Exploring the Role of Punctuation in Parsing Natural Text." In *Proceedings of the 15th International Conference on Computational Linguistics (COLING94)*, Vol. 1, pp. 421–425. Also published as http://xxx.lanl.gov/abs/cmp-lg/9505024.

加藤恒昭 (1995). "一般化弧を用いた A* 探索による非文の解析." 情報処理学会論文誌, **36** (10), 2343–2352.

加藤輝政, 小川清, 佐良木昌 (1997). "英語複文の構文解析と編集、その論理と方法." 研究報告 NL120-10, 情報処理学会.

金淵培 江原暉将 (1994). "日英機械翻訳のための日本語長文自動短文分割と主語の補完." 情報処理学会論文誌, **35** (6), 1018–1028.

日本電子工業振興協会 (1995). "機械翻訳システム評価基準—品質評価用テストセット—." 95–計–17.

Lewis, D. D. (1997). "Reuters-21578 Text Categorization Test Collection, Distribution 1.0." http://www.research.att.com/~lewis/reuters21578.html.

松本裕治 今一修 (1994). "頑健な自然言語処理の研究動向と課題." 研究報告 SLP1-2, 情報処理学会.

Mellish, C. S. (1989). "Some Chart-Based Techniques for Parsing Ill-Formed Input." In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL89)*, pp. 102–109.

成田一 (1994). "連体修飾節の構造特性と言語処理—日本語らしい表現の機械翻訳と応用技術—." 田窪行則（編）, 日本語の名詞修飾表現, pp. 67–126. くろしお出版.

佐川雄二, 大西昇, 杉江昇 (1994). "自己修復を含む日本語不適格文の分析とその計算機による理解手法に関する考察." 情報処理学会論文誌, **35** (1), 46–52.

白井諭, 池原悟, 河岡司, 中村行宏 (1995). "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, **36** (1), 12–21.

Shirai, S., Ikehara, S., Yokoo, A., and Ooyama, Y. (1998). "Automatic Rewriting Method for Internal Expressions in Japanese to English MT and Its Effects." In *Proceedings of the 2nd International Workshop on Controlled Language Applications*, pp. 62–75.

武田紀子 (1995). "英日機械翻訳システムにおける挿入句の処理." 研究報告 NL110-10, 情報処理学会.

富田春生 (1994). 英字新聞の読み方. 連合出版.

上野田守 布施敏夫 (1978). 新聞英語. 朝日実務英語シリーズ. 朝日出版社.
山口昌也, 乾信雄, 小谷善行, 西村恕彦 (1998). “前編集結果を利用した前編集自動化規則の獲得.” 情報処理学会論文誌, **39** (1), 17–28.
安井稔 (1982). 英文法総覧. 開拓社.

付録

表 6　形態素属性を調べる手続きの一覧

手続き名	説明
word(pos, w)	入力文中で位置 pos に単語 w が存在すれば真を，さもなければ偽を返す．pos には，p(着目語の位置)，$p+n(-3 \leq n \leq 5)$，com_fwd_$m+n(m \leq 3)$，com_bwd_$m+n(m \leq 3)$，0(文頭)，\$(文末) のいずれかが記述される．
word(pos_1, pos_2, w)	入力文中の区間 $[pos_1, pos_2]$ に単語 w が存在すれば真，さもなければ偽.
word_class(pos, wc)	位置 pos に語種が wc である語が存在すれば真，さもなければ偽．wc としては，adjective, noun, singular_noun, plural_noun, verb, past_participle など 59 種類が記述されうる．
word_class(pos_1, pos_2, wc)	区間 $[pos_1, pos_2]$ に語種が wc である語が存在すれば真，さもなければ偽.
sem_feat(pos, sf)	位置 pos に意味標識が sf である語が存在すれば真，さもなければ偽．sf としては，HUMAN，TIME，PLACE など 40 種類が記述されうる．
sem_feat(pos_1, pos_2, sf)	区間 $[pos_1, pos_2]$ に意味標識が sf である語が存在すれば真，さもなければ偽.
verb_pat(pos, vp)	位置 pos に動詞型が vp である語が存在すれば真，さもなければ偽．vp としては，Hornby の分類 (Hornby 1977) を拡張した 60 種類が記述されうる．
verb_pat(pos_1, pos_2, vp)	区間 $[pos_1, pos_2]$ に動詞型が vp である語が存在すれば真，さもなければ偽.
adj_pat(pos, ap)	位置 pos に形容詞型が ap である語が存在すれば真，さもなければ偽．ap としては，Hornby の分類を拡張した 9 種類が記述されうる．
adj_pat(pos_1, pos_2, ap)	区間 $[pos_1, pos_2]$ に形容詞型が ap である語が存在すれば真，さもなければ偽.
idiom(pos)	位置 pos の語が慣用句の構成要素ならば真，さもなければ偽.

略歴

吉見 毅彦:　1987 年電気通信大学大学院計算機科学専攻修士課程修了．1987 年よりシャープ (株) にて機械翻訳システムの研究開発に従事．1999 年神戸大学大学院自然科学研究科博士課程修了．

佐田 いち子:　1984 年北九州大学文学部英文学科卒業．同年シャープ (株) に入社．現在，同社情報システム事業本部ソリューション事業推進センター ソフト開発部係長．1985 年より機械翻訳システムの研究開発に従事．

福持 陽士:　1982 年インディアナ大学言語学部応用言語学科修士課程修了．翌年，シャープ (株) に入社．現在，情報システム事業本部ソリューション事業推進センター ソフト開発部副参事．機械翻訳システムの研究開発に従事．

(1999 年 12 月 6 日 受付)
(2000 年 1 月 17 日 再受付)
(2000 年 6 月 30 日 採録)