社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

格助詞およびその相当表現のパターン翻訳の試み

吉田　大蔵† 　徳久　雅人†† 　村上　仁一†† 　池原　悟††

† 鳥取大学工学部知能情報工学科
†† 鳥取大学大学院工学研究科情報エレクトロニクス専攻
〒 680–8552 鳥取県鳥取市湖山町南 4–101
E-mail: †s062062@ike.tottori-u.ac.jp, ††tokuhisa@ike.tottori-u.ac.jp

あらまし　本稿は，非線形言語モデルに基づく日英機械翻訳における格要素の翻訳について報告する．このモデルによる日英翻訳では，入力文の表現構造の意味を日文パターンで捉えるのだが，日文パターンから見て任意の格要素と判断された格要素は，対訳英文パターンでは翻訳されないという問題がある．そこで，本稿では，格要素と前置詞句のパターン対で構成する「格要素パターン辞書」を構築することを目的とする．既に構築された日文英文パターン対には格要素と前置詞句の対応が多く含まれていることに着目し，格要素と前置詞句のパターン対を抽出する．約 22 万件の重文複文文型パターン辞書から抽出した結果，925 件のパターン対が得られた．日本語語彙大系と組み合わせて，単文の日英翻訳実験を行ったところ，日本語格要素パターンによる表現に対する網羅性は問題が無く，英語前置詞句パターンによる意味的なカバー率が 79%であることが確認された．

キーワード　非線形言語モデル，格要素，前置詞句，パターン対，パターンベース翻訳，パターン辞書

A trial of pattern based machine translation for postpositional phrases

Taizo YOSHIDA†, Masato TOKUHISA††, Jin'ichi MURAKAMI††, and Satoru IKEHARA††

† Department of Information and Knowledge Engineering, Faculty of Engineering, Tottori University
†† Department of Information and Electronics, Graduate School of Engineering, Tottori University
4–101, Koyama-Minami, Tottori, 680–8552, Japan
E-mail: †s062062@ike.tottori-u.ac.jp, ††tokuhisa@ike.tottori-u.ac.jp

Abstract　This paper reports preliminary results of Japanese-English machine translation for postpositional phrases based on non-compositional language model. According to this model, the meaning of the input sentence is captured by sentence pattern. However, the postpositional phrases which are decided as compositional by the pattern can not be translated by the pattern. In order to translate such phrases, we construct "postpositional phrase pattern dictionary," which consists of the pairs of Japanese postpositional phrase pattern and English prepositional phrase pattern. At first, we extract the pairs from the dictionary of complex/compound Japanese-English sentence patterns, because this dictionary contains many pairs of post/pre-positional phrases described by letters and variables. As the results, 925 pairs of post/pre-positional phrase patterns are extracted. Next, in the experiment on Japanese-English machine translation of simple sentences, we confirmed that the Japanese postpositions are perfectly covered and the 79% of the English prepositions are covered and potentially translated by our dictionary.
Key words　non-compositional language model, case element, postpositional phrase, prepositional phrase, pattern pair, pattern based machine translating, pattern dictionary

1. はじめに

本稿は，非線形言語モデルに基づく日英機械翻訳 [1] における格要素の翻訳について現状を報告する．

非線形言語モデルは，まず大局的に，入力文の表現構造の意味を文単位のパターンで捉え，次に局所的に，そのパターンで定められる線形部分の意味を節，句，単語等を単位とするパターンで捉えるというモデルである．単文のパターンについては，日本語語彙大系で示されており [2]，重文・複文のパターンについては，「鳥バンク」で公開されている [3].

ここで，格要素は大きく 2 種類がある．1 つは文構造の意味を定める上で必須のもの，もう 1 つは任意のものである．必須のものとは格要素の内容が変わると文全体の解釈が変わるもので，たとえば,「小遣いを貰う」と「嫁を貰う」ではヲ格の内容の違いにより,「所有的移動」と「相対関係」という意味で違いが生じる．任意のものとは，その格要素の解釈とその格要素以外の解釈を合成して文全体の解釈とできるもので，たとえば,「バスで学校へ行く」では「学校へ行く」という意味に「バスで」という「手段」の意味を線形的に加えることで，文全体が解釈できる．

日本語語彙大系においては，必須の格要素が洗練されて示されているが，任意の格要素が体系的には示されいない．鳥バンクにおいては，約 12 万件の日英文対について，線形部分と非線形部分の識別によりパターン化が行われ，任意の格要素の存在可能な位置が離散記号「/c」で示されたが，やはり，任意の格要素を陽に示すことはできていない．

こうした状況では，パターン翻訳を行う際，入力文に文パターンが適合しても，任意の格要素の翻訳が出来ないという問題がある．その解決には，任意の格要素を解釈するためのパターン辞書を新たに作成する必要がある．

一般的に，日英翻訳において任意の（=必須でない）日本語格要素は英語前置詞句で翻訳することが適当であるとすると，先行研究に見られるアライメント手法で格要素と前置詞句の対を抽出することが，解決方法の 1 つと言える [5], [6]．しかし，本稿では，鳥バンクにおける重文・複文の文型パターン辞書において，既に単語・句のアライメントがマークアップされていることに着目し，この辞書から格要素と前置詞句の対を抽出する方法を試みる．

非線形言語モデルに基づく格要素の扱いにおいて，(1) 格要素と前置詞句の対が任意のものとして使用可能かどうか，(2) 任意であるならばその解釈をどうするか（たとえば,「バスで」が「手段」と解釈するように）という点を考えなければならない．このうち本稿では，格要素と前置詞句の対の収集の試行，および，収集した全てを任意格要素と仮定して使用した翻訳の問題分析までを行うことを目的とする．

2. 格要素パターン辞書の作成

本稿のパターンは, [3] の記述仕様に従い，変数，字面，関数，および，記号で記述する．日本語の格要素のパターンを「格要素パターン」，英語の前置詞句のパターンを「前置詞句パターン」と呼ぶことにする．さらに，格要素パターンと前置詞句パターンの対のことを単に「格要素パターン対」と呼び，格要素パターン対の集合を「格要素パターン辞書」と呼ぶことにする．本章では，格要素パターン辞書の作成方法を示す．

2.1 扱う格助詞と前置詞

今回使用する日本語格助詞は, [7] より,「が」を除く「を，に，と，で，へ，まで，より，から」とする．また，これらを含み，格助詞と同じような働きをする語句を格助詞相当句とする．格助詞相当句については，格要素パターン抽出時に作られるので，あらかじめ準備はしない．一方，英語前置詞に関しては，between を除く一般的な前置詞 90 件，および，その相当句 45 件を使用する．

格助詞と前置詞のパターンの関係を図 1 に示す．点線で囲まれた部分は今回使用する知識であり，囲まれていない格助詞相当句は，抽出時に作られることを示している．

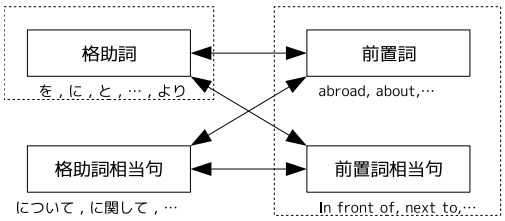

図 1　格助詞と前置詞の関係

2.2 鳥バンク

格要素パターン辞書を作成するにあたって,「鳥バンク」のデータを利用する.「鳥バンク」の「日本語表現意味辞書（重文複文編）」には，日本語の重文・複文とその対訳英文の対を約 12 万文対，および，日英で対訳となる単語・句・節が変数化された「意味類型パターン（22.7 万件）」が収録されている．そのパターンの例を図 2 に示す.

日文: 列車で行くほうが安全だろう。
英文: It would be safer to go by train.

日文パターン:
/ytcfk$N1$(OR:列車,NI:988,IM:13690) で/cf$V2$(OR:行く,NY:0506,NY:1500,NY:1801,NY:1804,NY:2001,NY:2301,NY:2302,NY:2303,NY:2903,NY:3201,IY:6970,IY:8410)^rentai/f ほうが/cf$AJV3$(OR:安全だろ,NY:0506,IY:A610).you。

英文パターン:
It would be $AJ3$(OR:safer)^er to $V2$(OR:go)^base by $N1$(OR:train).

図 2　日本語表現意味辞書における日英パターン対の例

図 2 では，日文の「列車」,「行く」,「安全」 という単語が日文パターンにおいてそれぞれ $N1$，$V2$，$AJV3$ という変数に変換されいるのがわかる．また，英文の「safer」,「go」,「train」が同様に $AJ3$，$V2$，$N1$ に変換されているのがわかる．さらに，それらの変数の直後に記述されている「(…)」は，その変数の用例，もしくは意味属性による制約を示している．

2.3 作成方法

「日本語表現意味辞書（重文複文編）」の日文英文パターン対から格要素パターン対を機械的に抽出する手順は次のとおりである．

（1） 日文パターンより格助詞を検索する．

（2） 格助詞の前後に付属語があるならば，格助詞相当句の一部とみなして記憶する．

（3） 格助詞または格助詞相当句の直前に名詞変数があるならば，その変数を記憶する．

（4） (2) と (3) より変数から格助詞もしくは格助詞相当句の終わりまでを格要素パターンとみなす．

（5） 英文パターンより，同一変数を検索する．

（6） (5) の変数の直前に前置詞または前置詞相当句があるならば，それと変数をあわせて前置詞句パターンとみなす．

（7） (4) の格要素パターンと (6) の前置詞句パターンを対にすることで，格要素パターン対を得る．

ここで，検索に失敗したり，変数や前置詞等が存在しない場合，格要素パターン対は得られない．また，変数を扱う際，変数の直後の制約を変数とともに抽出し，パターンに組み込むこととする．制約は「OR:...」より字面レベルでの条件記述がなされている．そこで，制約付きの格要素パターン対のことを「用例レベル格要素パターン対」と呼び，制約を取り除いた格要素パターン対のことを「文法レベル格要素パターン対」と呼ぶ．

2.4 抽 出 例

たとえば，図 2 の日文パターンにおいて，「で」の直前に変数 $N1$ が存在し，英文パターンの変数 $N1$ の直前に「by」が存在することから，これを格要素パターン対として取得する．この結果，下記の格要素パターン対が得られる．

格要素パターン対:
N_k (OR:列車,NI:988,IM:13690) で　by N_k (OR:train)

同様にして得られる格要素パターン対のいくつかを図 3 に示す．ここでは用例レベルの格要素パターン対が列挙されている．同一の格助詞と同一の前置詞から成る対がみられる．したがって文法レベルのパターン対は用例レベルに比べて種類数が少ない．

N_k (OR:教会,NI:377,NI:455,IM:11420,IM:12150) で　at N_k (OR:a church)
N_k (OR:金利,NI:1198,NI:2596,IM:14A50,IM:16740)で　at N_k (OR:rates)
N_k (OR:浜辺,NI:2667,NI:490,IM:12340,IM:168C0) で at N_k (OR:the beach)

N_k (OR:バス,NI:1056,NI:1349,NI:2354,NI:868,NI:988,IM:13660,IM:13690,IM:14230,IM:15140,IM:1541C) で　by N_k (OR:bus)
N_k (OR:電話,NI:1147,NI:1548,NI:970,IM:13680,IM:14800,IM:151B0)で　by N_k (OR:phone)

N_k (OR:雪,NI:2365,NI:744,IM:13540,IM:15424)で　because of N_k (OR:the snow)
N_k (OR:病気,NI:2416,NI:2419,IM:15450)で　because of N_k (OR:illness)

図 3　格要素パターン辞書の一部

2.5 作 成 結 果

格要素パターン対，および，格助詞相当句を抽出した結果を表 1 に示す．

表 1　抽 出 結 果

抽出結果	抽出件数
用例レベルでの格要素パターン対総数	10,783
文法レベルでの格要素パターン対総数	925
格助詞相当句数	266

また，得られた格助詞相当句の例を図 4 に示す．

からの　からは　からも　ごとに　だけで　だけに　では　でも　として　どおりに　と同時に　について　にとっては　に関して　に沿って　に対して　という　として　ときたら　としての　を通じて　なしで　…

図 4　格助詞相当句の例

3. 機械翻訳の試み

本章では，日本語語彙大系の文パターン辞書と本稿の格要素パターン辞書を用いた翻訳を試みる．

3.1 文パターン

日本語語彙大系には，結合価文法に基づく文型パターン（以下，文パターンと呼ぶ）が定義されている．パターンに記述された変数には意味属性による制約条件が付けられている．また，パターンには英文パターンも記述されている．具体例を図 5 に示す．

日文パターン:
$N1$ が $N2$ に/へ/まで $N3$ から/より 行く
制約条件:
$N1$(3 主体 986 乗り物 535 動物) $N2$(-418 道路 -419 鉄道 1057 創作物 (その他) 388 場所 2610 場) $N3$(-418 道路 -419 鉄道 388 場所 2610 場)
英文パターン:
$N1$ go from $N3$ to $N2$

図 5　日本語語彙大系データの 1 レコード

3.2 翻 訳 方 法

図 6 を参照しながら，日英翻訳の手順を説明する．

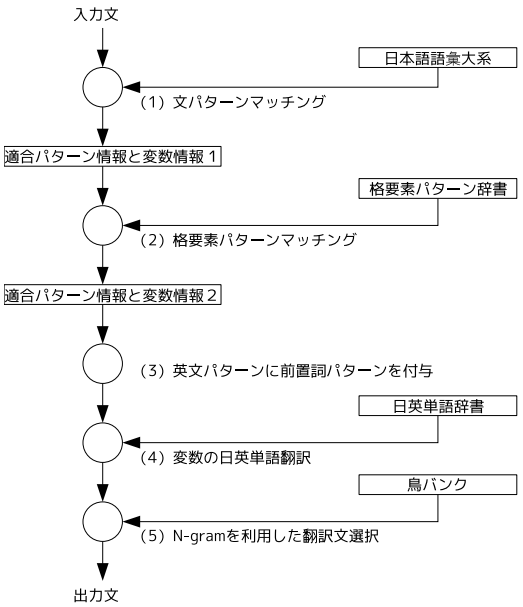

図 6　翻訳処理の流れ

（1） 日本語の入力文に対して，まずは日文パターンを照合し，次に意味属性制約を検査する．それにより適合した情報が「適合パターン情報と変数情報 1 」である．これは，適合した日文パターン，それに対応する英文パターン，および，文パターンの変数に対応する単語の情報で構成する．

（2） (1) で補えなかった格要素を格要素パターン辞書の格助詞パターンと照合し，ここでも変数の意味属性制約を検査する．この結果（適合した格要素パターン，それに対応する前置詞句パターン，および，変数の情報）を (1) の出力に付与したものが「適合パターン情報と変数情報 2 」である．

（3） (1) の英文パターンに，(2) の前置詞句パターンを付与する．このときの前置詞句パターンで，格要素パターン辞書作成時に，抽出元となった英文パターンの先頭で使われていたものは (1) の英文パターンの先頭に，それ以外で使われていたものは後方にそれぞれ付与する．それらの判断は前置詞句パターンの先頭が大文字か小文字で行う．なお，英文パターンで入力文に対して必要のない前置詞句部分は削除する．

（4） (3) によって作られた英文パターンの変数に対応する日本語を日英単語翻訳する．ここで，変数は名詞や形容詞であるので，単なる辞書引きで実現できる．このとき，複数の候補がある場合，全てを翻訳候補として登録する．

（5） 入力文に対する複数の翻訳候補（パターンと英単語の両方に曖昧性がある）に対して，英単語の tri-gram で文のスコアを算出する．このスコアにより，出力文を 1 つにする．

3.3 翻 訳 例

翻訳の具体例を示す.「学生がバスで学校に行く。」という入力文を文パターンに照合した様子を以下に示す.

複数の適合パターンの中の 1 つを図 7 に示している．この情報は図 6 の「適合パターン情報と変数情報 1 」に対応している．

適合パターン情報：
$N1$ が $N2$ に/へ/まで $N3$ から/より 行く。
$N1$(3 主体 986 乗り物 535 動物) $N2$(-418 道路 -419 鉄道 1057 創作物 (その他) 388 場所 2610 場) $N3$(-418 道路 -419 鉄道 388 場所 2610 場)
$N1$ go from $N3$ to $N2$

変数情報: $N1$ = 学生 , $N2$ = 学校
補えなかった格要素:バスで

図 7 適合パターン情報と変数情報 1

補えなかった格要素を格要素パターン辞書と照合する例を図 8 に示す．この情報を「適合パターン情報と変数情報 1 」に付与したものが，図 6 の「適合パターン情報と変数情報 2 」に対応している．

適合パターン情報：
N_k で in N_k (OR:自動車,NI:988,IM:13690)
N_k で on N_k (OR:電車,NI:988,IM:13690)
N_k で by N_k (OR:バス,NI:1056,NI:1349,NI:2354,NI:868,NI:988, IM:13660, IM:13690,IM:14230,IM:15140,IM:1541C)
…

図 8 「適合パターン情報と変数情報 2」の元になる格要素パターン候補群

図 8 の情報をもとにして，前置詞句を英文パターンに付与した結果を図 9 に示す．図 8 では,「at N_k」,「by N_k」のパターンの前置詞の先頭は小文字なので，図 9 の新規英文パターンでは後方に付与されている．また，図 7 の英文パターンで存在した「from $N3$」は，入力文に対して必要のない前置詞部分であるので削除している．

新規英語パターン:
$N1$ go to $N2$ in $N3$
$N1$ go to $N2$ on $N3$
$N1$ go to $N2$ by $N3$
…

図 9 新規英語パターンの例

これらによって新しく作成された英文パターンの変数に対応する日本語の日英単語翻訳を行う．そして，複数の翻訳候補の中から tri-gram によって選出を行う．日英翻訳した例を図 10 に示す．任意格要素に対する前置詞句パターンの候補が複数ある中から，tri-gram により,「by」がもっとも高いスコアとなり，出力文のパターンに選ばれている．

入力文：学生がバスで学校に行く。

適合パターン情報:
$N1$ が $N2$ に/へ/まで $N3$ から/より 行く。
$N1$ go from $N3$ to $N2$
$N1$(3 主体 986 乗り物 535 動物)
$N2$(-418 道路 -419 鉄道 1057 創作物 (その他) 388 場所 2610 場)
$N3$(-418 道路 -419 鉄道 388 場所 2610 場)
任意格要素:バスで

前置詞句の候補と新英文パターンによる翻訳時の文の N-gram の値:

by N_k:	-48.690
of N_k:	-52.434
for N_k:	-58.266
to N_k:	-58.500
in N_k:	-58.736
on N_k:	-58.959
with N_k:	-59.365
at N_k:	-59.652
but N_k:	-61.444
After N_k:	-68.572
With N_k:	-68.863
like N_k:	-69.354

出力された翻訳文:Students go to school by bus

図 10 翻 訳 例

4. 評価実験

入力文の任意格要素に対する格要素パターン辞書の効果を評価する．

4.1 評価方法

4.1.1 実験文の条件

実験には，辞書の例文より抽出した単文コーパス [4] から，格助詞を含む簡単な単文 100 文に対して翻訳を行う．今回，使用する文を説明する．

- 能動態である文
- 連用修飾語を含まない文
- 形態素解析に成功した文
- 日本語語彙大系の結合価パターンと適合する文

本稿で評価の対象とする文は，正解英語文が前置詞を使用した文であることと，入力文に対して日本語語彙大系の文パターンだけでは補えず，任意格要素の翻訳が必要である文である．以下このような文を任意格要素翻訳必要文とする．

以上のように，はじめから評価対象の文が決まらないのは，任意格要素翻訳必要文の判定基準が日本語語彙大系のパターンにあるためである．

4.1.2 評価の分類

次に，評価する基準を説明する．

◎　…出力文において，正解文と同じ前置詞が存在したもの

○　…◎以外で，出力文の候補において，正解の前置詞が存在したもの

△　…入力文の任意格要素に対して，格要素パターンが適合したが，正解の前置詞句パターンが出力文候補に含まれていないもの

×　…入力文の任意格要素に対して，格要素パターンが使われていないもの

4.2 実験結果

実験を行った結果を表 2，3 に示す．まず，表 2 より，任意格要素の翻訳が必要な入力文は，47 件であった．

次に，任意格要素翻訳必要文 47 件に対して評価を行った結果が表 3 である．正解の前置詞句パターンが翻訳候補に含まれたものは，◎と○を合計して，37 件である．これは，任意格要素に対する格要素パターン辞書の意味的なカバー率が，79% (37/47) であると解釈できる．一方，×が 0 件であることについては，日本語の表現レベルでのカバー率が，100%であると解釈できる．したがって，格要素パターン辞書は，基本的には成功理に作成できたと言える．

表 2　入力文における任意格要素翻訳必要文数

全入力文数	任意格要素翻訳必要文数
100	47

表 3　4.1.2 節に基づく出力文の評価の内訳

◎	16 件
○	21 件
△	10 件
×	0 件

5. 考　　察

任意格要素翻訳必要文 47 件に対して，◎は 16 件であることから，正解率は 34%である．これは，前置詞句の翻訳候補の選択能力が原因である．すなわち，格要素パターン辞書の運用上の問題である．運用をよりよくするために，○と△の事例を分析する．

5.1 ○ の 例

○の例を図 11 に示す．前置詞句パターンの候補には，正解文の前置詞「with」が含まれているが，翻訳時の文の tri-gram が「in」や「of」より低いことから，正しく前置詞句パターンが選ばれていなかった．今後，より高度なスコア付けが必要である．

入力文：彼は優秀な成績で大学を卒業しました。

適合パターン情報:
$N1$ が $N2$ を 卒業する
$N1$ graduate from $N2$
$N1$(4 人) $N2$(362 組織 1002 抽象物 (精神) 1237 精神)
任意格要素:優秀な成績で

前置詞句の候補と新英文パターンによる翻訳時の文の N-gram の値:

in N_k:	-58.809
of N_k:	-60.119
with N_k:	-60.555
for N_k:	-60.691
on N_k:	-61.087
but N_k:	-63.332
like N_k:	-73.121
at N_k:	-73.390
WithN_k:	-76.320

出力された翻訳文:He graduate from university in excellent results
正解文:He graduated from college with excellent records .

図 11　正解格要素パターンが選ばれなかった例

5.2 △ の 例

△の例を図 12 に示す．補うべき前置詞句の候補に正しい前置詞句パターンがなかった．格要素パターン辞書の用例が不足していたことが原因であるのだが，一方で，候補の過度の削減も原因である．これを抑制するために，今後，意味属性制約の緩和が対策として考えられる．

入力文：ノートが消しゴムで破れた。

適合パターン情報:
$N1$ が 破れる
$N1$ be torn
$N1$(533 具体物)
任意格要素:消しゴムで

前置詞句の候補と新英文パターンによる翻訳時の文の N-gram の値:
With N_k: -75.260
with N_k: -81.118
for N_k: -81.384
like N_k: -81.517
in N_k: -81.629
but N_k: -82.926

出力された翻訳文:With an eraser notebook be torn
正解文:A page of the notebook was torn by an eraser .

図 12　正解前置詞句が出力文候補に含まれていない例

6. おわりに

本稿では，非線形言語モデルに基づく日英機械翻訳における任意の（必須でない）格要素の翻訳のための格要素パターン辞書の構築を行ない，その翻訳方法を示した．まず，格要素と前置詞句の対の収集を約 22 万件の重文複文文型パターン辞書から行った結果，用例レベルでの格要素パターン対を 10,783 件，文法レベルでの格要素パターン対を 925 件が得られた．同時に，格助詞相当句を 266 件が得られた．次に，格要素パターン辞書を用いて単文の日英機械翻訳を実験したところ，格要素パターン辞書の任意格要素に対する意味的なカバー率が 79%，および，日本語の表現レベルでのカバー率が 100%という結果を得た．したがって，格要素パターン辞書は，基本的には成功理に作成できたと言える．今後の課題は，格要素パターン辞書の運用の方法を精緻することで翻訳性能を高めること，一方で，任意格要素に対して意味分類を付与し，意味解析に役立てることである．

文　献

[1] 池原悟: 非線形言語モデルによる自然言語処理, 岩波書店, 2009.
[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: 日本語語彙大系, 岩波書店, 1997.
[3] 鳥バンク, 日本語表現意味辞書 - 重文複文編 -, http://unicorn.ike.tottori-u.ac.jp/toribank
[4] 西山七絵, 村上仁一, 徳久雅人, 池原悟: 単文文型パターン辞書の構築, 言語処理学会第 11 回年次大会, pp.372-375, 2005.
[5] Yamamoto,K., Matsumoto,Y.: Extracting Translation Knowledge from Paralel Corpora, M.Carl and A.Way(eds), Recent Advances in Example-based Machine Translation, pp.365-395, Kluwer, 2003.
[6] Och, F.J, Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, Vol.29, No.1, pp.19-51, 2003.
[7] 益岡隆志, 田窪行則: 基礎日本語文法, くろしお出版, 1989.
[8] 和田吉剛: アットウィル総合英語, 美誠社, 2003.