自然言語処理 49－7
（1985. 5. 24）

Muプロジェクトにおける日英翻訳実験の形態素処理の結果とその評価

坂本義行　　中村順一　辻井潤一　長尾　真
（電子技術総合研究所）　（京都大学工学部）

概　要

Muプロジェクト＊における言語処理システムの日英翻訳の大規模な実験を行っている．ＪＩＣＳＴ抄録　電気編3,000文について，３回の実験を行い，各回ごとに実験結果を評価検討し，翻訳ソフトウェア，文法，辞書の改良と拡張を行っている．本論文は，第１回翻訳実験（５９年１１月終了）と第２回翻訳実験（６０年１月終了）の実験結果について報告する．

本実験は，日本語形態素解析，日本語構文解析，日英変換，英語構文生成，英語形態素生成に分割されて処理されている．構文解析，変換，生成については，別に報告がなされているので，本稿では，ＲＩＰＳ（工業技術院筑波情報計算センター）で行った実験の全体の様子と形態素解析，生成についての結果および評価について述べる．

１．はしがき

Muプロジェクトは，１９８２年度より基本設計，詳細設計，基本部分の作成と３年間にわたって研究開発が進められてきた．

日英翻訳の中心部分をなす言語処理システムは前述のように日本語形態素解析，日本語構文解析，日英変換，英語構文解析，および英語形態素生成の各部分に分割して開発した．形態素解析と生成部分はＵＴＩＬＩＳＰで直接記述する形で電子技術総合研究所が担当して開発し，構文解析，変換生成の部分は本プロジェクトのために開発されたＧＲＡＤＥを用いて記述する形をとり京都大学で開発した．1,000文のサンプル文を対象に，昨年までに文法として解析変換，生成の各規則の作成および日本語辞書，日英変換辞書，英語生成辞書の作成を行った．本年は，3,000文に対象を広げ，各部分の改良と拡張を行ってきた．各部分の完成度をみるために昨年秋からＪＩＣＳＴ抄録　電気編のタイトルと抄録文の大規模な翻訳実験を開始した．

同一のテキストに対し３回の実験を繰り返すことを計画し，第１図に示すように，１回の実験を行うごとに各文についてその処理速度，翻訳各段階での処理結果，えられた翻訳文と原文から翻訳の質に関する評価を外部に委託して行った．この実験結果をフィードバックして，主に文法ルール，辞書の項目の改良，拡張およびプログラムの細部のチェックを行ってきた．

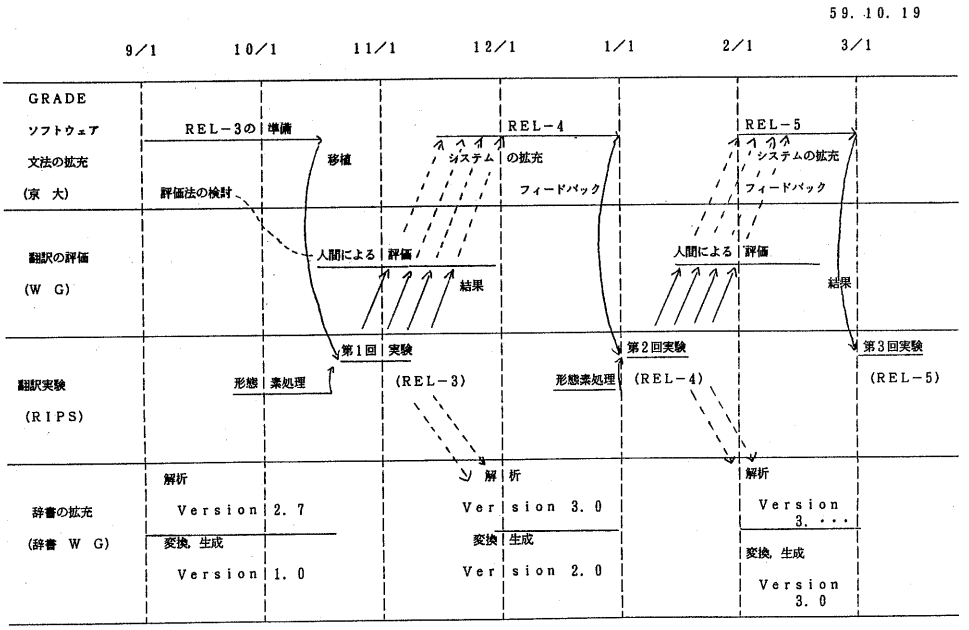
２．実験の環境

（１）計算機システム

ＦＡＣＯＭ　Ｍ３８０システム

リージョンサイズ　４ＭＢ～８ＭＢ

＊本研究は国の科学技術振興調整費による「日英科学技術文献の速報システムに間する研究」の一部として行ったものである．その遂行のために言語処理システム作業分科会を組織し，その審議，指導のもとに研究を進めている．



第１図　５９年度日英翻訳(3000文) のテストスケジュール

ＢＡＴＣＨ処理　Ｂクラス

（２）ソフトウェア

ＯＳ：ＯＳ　ＩＶ／Ｆ４　ＭＳＰ

プログラミング言語：ＵＴＩＬＩＳＰの漢字版，Garbage Collectionをディスク上で行える機能ならびに，ＶＳＡＭアクセス機能を有する．

（３）ファイル環境

翻訳用プログラム，入出力テキスト，文法規則，辞書類，翻訳用ユーティリティ汎用ファイル上に記憶するために１００ＭＢ以上と翻訳処理の各段階での結果と辞書用として２００ＭＢ以上のＶＳＡＭファイルを必要とする．

３．翻訳実験

実験は前節の実験環境のもとにＲＩＰＳ共同利用室の日本語端末からＢＡＴＣＨ処理にて行った．

（１）翻訳対象文

ＪＩＣＳＴ抄録　電気編　約3,000文(第25巻6号,1号(半月)分),文献数 745中のタイトルおよび抄録文を文単位で翻訳を行った．

（２）翻訳用辞書（Ｖ２．７）

日本語解析辞書

日英変換辞書

英語生成辞書

用言　約3,000語,体言 約13,000語,付属語 400語

（３）文法規則

日本語解析文法ルール；SN=85,SG=444,RR=1723

日英変換文法ルール;SN=33,SG=716,RR=894

英語生成文法ルール；SN=54, SG=301, RR=629

SN=Subgrammar Network

SG=Subgrammar

RR=Rewriting Rule

（４）言語処理ソフトウェア

日本語形態素解析プログラム；3,000 行

ＧＲＡＤＥ（構文解析，日英変換，英語生成）；15,000行

英語形態素生成プログラム；1,000 行

（５）処理単位

第１回の実験では，形態素解析（ＭＡ），構文解析（ＳＡ），変換（ＴＲ），生成（Ｇ，構文生成と形態素生成を一括して）の４段階の処理に分割して行った．ＭＡは処理速度が速いので３０文献を処理の単位とし，ＳＡ，ＴＲ，Ｇは３文献（約１５文）を処理単位とした．

第１表　第１回実験の各段階のＣＰＵ処理時間

各処理段階	処理単位当り（ｓｅｃ）	処理回数	合計時間（ｍｉｎ）
ＭＡ	２０	２５	８．３
ＳＡ	２１０	２５０	８７５
ＴＲ	３	２５０	１２．５
Ｇ	９０	２５０	３７５

第２回の実験では，ＭＡは第１回と同様に３０文献を単位にとったが，ＳＡ，ＴＲ，Ｇは途中の処理段階で処理が不成功の場合にもその文の処理を最後まで続行できるようにシステムを改良して一文を単位として一括処理による効率化を計った．

各処理段階での出力結果をＶＳＡＭファイル上に記憶し，次の処理への入力データとした．これにより処理に問題が発生した場合等に再処理を効率的に行えるようにした．

（６）処理に要した時間

第１回実験

第１表に示すようにＣＰＵ時間として，ＭＡ，ＳＡ，ＴＲ，Ｇの各段階での処理に時間を要した．実際にはバッチ処理で行っており，ＣＰＵで１分の実行に要するターンアラウンドタイムが，混雑時には１時間近くを要した．

3,000文全体の処理時間は約20時間をようした．従って実際の作業時間（ターンアラウンド）としては，週４日で２４時間連続処理を行っても１ケ月以上かかった．

第２回実験

この実験では各文に対して処理が成功した段階までの消費時間を計算し，その平均時間を第２表に示した．

処理時間は翻訳が成功した場合が最も少なく変換，生成，解析が失敗した順に時間

第２表　第２回実験の各段階のＣＰＵ処理時間

平均処理時間	ＳＡで失敗	ＴＲ失敗	Ｇ失敗	成功
ＣＰＵタイム（ｓｅｃ）	２６８	１９１	２３７	３８．６
ターンアラウンドタイム（ｈ：ｍ）	１：１３	１：１６	１：０６	１：０２

が増加し解析段階で失敗した場合成功した場合の７倍近い時間を消費している．成功した場合１文献当たり平均４文とすれば１文で費やす時間は１０秒以内である．

（７）翻訳実験の各処理段階の出力

第１図で示したようにＲＩＰＳで翻訳を行い，その結果が翻訳の評価グループ，言語処理システムの文法開発グループおよ

```
+PSANAPROG: E82060522_2_1
高速ＣＡＭＡＣデータ収集用の多用途マイクロコンピュータ。
/高速ＣＡＭＡＣデータ収集/用/の/多用途マイクロコンピュータ/。/

MESSAGE OF JAG EXECUTION.

  ALL CPU : 4012   MSEC. ( GC TIME : 788   MSEC.  GC : 5  )
TREE IS SAVED TO THE VSAM DATASET.
PROCESS(ANA) END.

START: 02/09/85 01:07:28, END: 02/09/85 01:07:59.
TIME: 4682 MS.(GCTIME: 788 MS., GCCOUNT: 5).
+GGG-JE-LOOP-2: PROCESS:A, KEY:E82060522_3_1:   START EXECUTION...
NOW TREE PRINTER MODE IS J.
+PSANAPROG: E82060522_3_1
ＡＭＤ２９００マイクロプロセッサと高速ＴＴＬ技術を用い，適当なインタフェース・ボードを選べば，ＣＡＭＡＣ構成機器となる。
/ＡＭＤ２９００マイクロプロセッサ/Ｐ と/高速だ/? ＴＴＬ/技術/を/用いる/,
/適当だ/インタフェース・ボード/を/選ぶ/ば/, /ＣＡＭＡＣ構成機器/Ｐ と/なる
/。/
(((選ぶ)))
(((成る)))
(((適当だ)))
"% % % % % ***_A_SYAKUDO.RR IS CALLED"
(((高速だ)))

MESSAGE OF JAG EXECUTION.

  ALL CPU : 24774  MSEC. ( GC TIME : 10457 MSEC.  GC : 58 )
TREE IS SAVED TO THE VSAM DATASET.
PROCESS(ANA) END.
```

第２図　日本語解析が成功した場合

```
"%%%%% J_***_DISAMB WAS CALLED,IT MAY BE"
"%%%%% AN ERROR"
"%%%%% PLEASE SHOW THE FOLLOWING"
"%%%%% TO YOUR GRAMMAR ADMINISTRATOR"
(((¥SYS¥ (PARA)))
 (((＄見出し語 (いる))
   (＄動詞活用型 (上一))
   (＄特殊活用形
      (＄未然形 ((＄形 "い") (＄後接情報 34)))
      (＄連用形 ((＄形 "い") (＄後接情報 7)))
      (＄連体形 ((＄形 "いる") (＄後接情報 10)))
      (＄終止形 ((＄形 "いる") (＄後接情報 9)))
      (＄仮定形 ((＄形 "いれ") (＄後接情報 33)))
      (＄命令形 ((＄形 "いろ") (＄後接情報 11)) ((＄形 "いよ") (＄後接情報 11))))
   (＄構文品詞 (動詞))
   (＄格パターン
      (V1 (＄アスペクト (状態))
          (＄態 (使役))
          (＄意志 (有))
          (＄格支配情報
             ((＄表層格 が) (＄深層格 主体) (＄名詞意味コード ＯＨ) (＄必須性 1))
             ((＄表層格 に) (＄深層格 場所) (＄名詞意味コード ＴＳ) (＄必須性 1)))))
   (＄動詞パターン (に が))
   (＄動詞表層格・深層格 (に__場所 が__主体))
   (＄連用中止 (意志動詞))
   (＄語尾 (る))
   (＄活用形 (連体形))))
 (((＄見出し語 (いる))
   (＄動詞活用型 (上一))
   (＄特殊活用形
      (＄未然形 ((＄形 "い") (＄後接情報 34)))
      (＄連用形 ((＄形 "い") (＄後接情報 7)))
      (＄連体形 ((＄形 "いる") (＄後接情報 10)))
      (＄終止形 ((＄形 "いる") (＄後接情報 9)))
      (＄仮定形 ((＄形 "いれ") (＄後接情報 33)))
```

第３－１図　日本語解析が不成功の場合の辞書情報

び辞書グループに渡される方法がとられた．そのため各処理段階の結果および誤りが発生した場合の状況リストを出力するようにした．以下に各段階での出力結果を示した．

```
?  <＊＊＊  ERROR  ERROR  ERROR      ＊> <>                    1
|--S   <?>  <>                                                2
|   |--S  <?>  <>                                            3
|   |   |--ADVP  <サイクル時間>  <>                           4
|   |   |   |--NP  <サイクル時間>  <>                         5
|   |   |   |   |--ADJP  <?>  <>                              6
|   |   |   |   |   |--N  <基本命令>  <>                      7
|   |   |   |   |   |
|   |   |   |   |   |--BKK  <NO>  <>                          8
|   |   |   |   |
|   |   |   |   |--N  <サイクル時間>  <>                      9
|   |   |   |
|   |   |   |--BFK  <HA>  <>                                  10
|   |   |
|   |   |--ADVP  <n s>  <>                                    11
|   |   |   |--N  <n s>  <>                                   12
|   |   |   |   |--?  <?>  <>                                 13
|   |   |   |      |--N  <1 5 0>  <>                          14
```

第３－２図　日本語解析が不成功の場合の解析木

（ａ）日本語解析

解析が成功した場合（第２図）

文法等のメッセージ，入力テキスト，形態素結果，用言リストおよび処理時間

```
((DOCUMENT_NO ("E 8 2 0 6 0 5 2 2")) (SENTENCE_TYPE (本文))
 (SENTENCE_NO (5)) (SUB_SENTENCE_NO (1)) (TRANSLATION_FLAG (T))
 (¥SYS¥ (ROOT)) (J_PASS_3 (YES)) (J_JAG (DEP))
 (J_LEX (＊＊＊  ERROR  ERROR  ERROR  ＊＊＊)) )

((J_PARA (T)) (J_RT (WV)) (J_INF (SYUUSI)) (J_CAT (S)) (J_STEP_1 (YES))
 (J_SENTENCE_RELATION (COMPOUND)) (J_DEEP_CASE (MAIN))
 (J_DEEP_ASPECT (TYOUJI)) (J_DEEP_TENSE (PRESENT)) )

((J_INF (RENYOU)) (J_VP (V2)) (J_RT_SS (T)) (J_SENTENCE_END (RENYOU))
 (J_CAT (S)) (J_RT_MARK (T)) (J_PASS_3 (YES)) (J_GCS1ED (T))
 (J_DEEP_CASE (BOTH)) (J_SS_IMI_PASS (T)) (J_RT (NV))
 (J_DEEP_ASPECT (JYOUTAI)) (J_DEEP_TENSE (PRESENT)) )

((J_SURFACE_BFK (HA)) (J_SEM (TAG)) (J_LEX (サイクル時間))
 (J_N (MFT)) (J_COUNT (0)) (J_TYPE (NP)) (J_CAT (ADVP)) (J_LEFT)
 (J_DEEP_CASE (TOKI)) (J_LIMIT (T)) (J_FREE (T)) (J_ADVP_HA (T))
 (J_TOPIC (T)) (J_DECIDE_CASE (YES)) )

((J_LEX (サイクル時間)) (J_N (MFT)) ($意味コード (TAG))
```

第３－３図　日本語解析が不成功の木のノード情報

```
MESSAGE OF PRE_MAIN_GENERATION EXECUTION.

  ALL CPU : 3644   MSEC. ( GC TIME : 1312  MSEC.  GC : 6  )
TREE IS SAVED TO THE VSAM DATASET.
PROCESS(GEN) END.

START: 02/09/85 01:52:19, END: 02/09/85 01:53:41.
TIME: 3700 MS.(GCTIME: 1312 MS., GCCOUNT: 6).
+GGG-JE-LOOP: START EMORG...
;;; @NP@
NO. 1: E82060522_2_1 (02/09/85, 02/09/85, 02/09/85)
高速ＣＡＭＡＣデータ収集用の多用途マイクロコンピュータ。

 Multipurpose microcomputers for the high - speed CAMAC data collection.

NO. 2: E82060522_3_1 (02/09/85, 02/09/85, 02/09/85)
ＡＭＤ２９００マイクロプロセッサと高速ＴＴＬ技術を用い，適当なインタフェース・ボードを選べば，ＣＡＭＡＣ構成機器となる。

 CAMAC component devices are provided by using AMD 2900 micro - processor and high-speed TTL technology when
appropriate interface boards are selected.

NO. 3: E82060522_4_1 (02/09/85, 02/09/85, 02/09/85)
これらでＣＡＭＡＣやＧＰＩＢの分岐構成を変化できる。

 The branch configuration of CAMAC and GPIB changes, and the formation is made by these.

NO. 4: E82060522_6_1 (02/09/85, 02/09/85, 02/09/85)
基本ソフトウエアはクロスアセンブラ，エミュレーション・プログラム，対話式デバッグ装置からなる。

 The basic software consists of cross assemblers, emulation programs and interactive debug devices.
```

第４図　英語生成結果のメッセージ

```
KEY                :AOUT                TROUT                GOUT
E82060522_2_1      :02/09/85 01:07:34   02/09/85 01:43:23    02/09/85 01:46:32
E82060522_3_1      :02/09/85 01:08:15   02/09/85 01:43:50    02/09/85 01:47:34
E82060522_4_1      :02/09/85 01:16:43   02/09/85 01:45:08    02/09/85 01:50:05
E82060522_5_1      :02/09/85 01:21:55X
E82060522_6_1      :02/09/85 01:38:50   02/09/85 01:46:05    02/09/85 01:52:20
-------------------
SUCCESS            :    4/    5 ( 80%)      4/    5 ( 80%)      4/    5 ( 80%)
ESCAPED            :    1/    5 ( 20%)      0/    5 (  0%)      0/    5 (  0%)
NIL

(QUIT)=>
```

第５図　翻訳処理結果のリスト

解析が不成功の場合（ｄｉｓａｍｂが発生した場合）（第３図）

動詞の辞書情報（第３－１図），解析木（第３－２図），木のノード情報（第３－３図）

（ｂ）日英変換（第３図）

未知語等に関するメッセージ

（ｃ）英語生成（第４図）

文法等のメッセージ，文単位の処理経過メッセージ，原文と訳文リスト

（ｄ）処理結果のリスト（第５図）

４．形態素解析の結果とその評価

（１）実験結果

入力テキスト3,000文について日本語形態素解析を行い，その出力結果に対して，ＭＡのみで判明する解析誤りを機械的に検出するため，次の２項目についてＫＷＩＣリストを出力するためのプログラムを作成した．

（ａ）ＭＡにおいて，単語間の接続関係が正常でないと判定された語について接続不可のＫＷＩＣリストを作成する．

第３表　形態素解析結果

実験	誤り項目	１－３００文献	３０１－６００文献	６０１－７４５文献
第１回	接続不可	５４	１８０	２８８
実験	未知語	７３３		１，５７７
第２回	接続不可	６	２８	９７
実験	未知語	１４０	１９９	－

注：個数は延べ語数である．
１－３００文献は，すでに数回の実験がなされている．
６０１－７４５文献中には未登録の名詞が多く含まれている．

（ｂ）ＭＡにおいて，辞書ならびに特殊記号，数式等のテーブルに未登録の語であると予想される語を未知語としてそのＫＷＩＣリストを作成する．その結果第３表のような数値がえられた．

この結果をみると接続不可も未知語も１回に比べ２回目ではその数が急激に減少していることがわかる．３回目では接続不可はほぼなくなるであろう．しかし６０１－７４５文献では名詞の辞書の追加作業が完成していないため非常に多数の未知語が残っている．

（２）接続不可の特徴

次に接続不可の特徴を第４表に示した．この表の中でＮｏ１－５までは辞書の問題である．ただし２回目ではＮｏ１の英字列とカナ文字列に対して字種による判別機能を処理システムに加えたのでその数は減少した．Ｎｏ９の誤りが３０個ほど出現した．

（３）問題事項と解決方法

解決方法として大きく分けて次の３種類がある．

（ａ）辞書に登録する．特に名詞以外はすべて登録する．

（ｂ）式や記号類はテーブルに登録しそのパターンマッチング機能を強化する．

（ｃ）ＭＡの処理機能を強化する．特に複合名詞や接辞等の未知語の処理

個々の問題事項について

数的処理

l個数，m個数

M６８００，ＡＮＡＰ－６

１０－３倍

化学式

特殊記号の処理

同じ雑誌７－，４５５（'８１）

…するか　：　コンジョィント

第６図　マルチパスのテスト例

入力の日本語文

低コスト，低不良率で高品質の製品が得られ溶剤が不要などの特徴がある．

解析結果

低コスト	低コスト
	低／コスト
低不良率	低不良率
	低／不良率
高品質	高品質
	高／品質
不要などの	不要な／どの
	不要／など／の

（４）形態素解析のマルチパス　モード

現在の実験システムでは，シングルパスモードで行っている．その理由として，マルチパス　モードで行うと以下のような多数の解析を出力する点および第１パス以外で正解をうるケースは非常にまれであるというテスト結果がえられたためである．その例を第５図に示す．

なお，マルチパスではあるが，制限事項を設けている．

マルチパスモードの制限事項

（ａ）２パス以降の処理では接続不可が発生したパスは採用しない．

（ｂ）分割結果の最上位文字列が等しいパスがえられたとき，最初にえられたもののみを出力する．

第４表　接続不可の特徴

	誤り項目	誤った例
1	名詞が未登録であるため	連用形転成名詞，英字，カナ文字，記号類
2	名詞以外が　〃	を識別，測定する．に代って，かどうか，のでなく，先駆ける
3	複合語が　〃	重イオン方式
4	異形語が　〃	製かん，（７０年）半ば，あゆみ
5	接尾辞が　〃	さ，性，内，なみ，式，あゆみ
6	単位が　〃	０．３９ＰＴ／Ｓ，ＭｅＶ
7	カッコの処理	温度（℃），非対角(off-diagonally)
8	読点との接続	明確化，である．　と調査，の内容について．誤差の検討について述べる．
9	入力テキストの誤り	…が主てあり，胴察する，人口貯水池，精能，方向遊離えの
10	その他	困難な複雑な計算，不要などの，正確適切な

第５表　数式処理の対象例

演算子	数式の例
＝	入射角θ＝４５°で　は１／λ＝１．５，ｄ／ｌ＝０．２，ｈ／λ＝３．５／ｌただしｌ＝格子　α／Ｐ０＝Ａｅｘｐ（－ＢＰ０／Ｅ）における定数Ａ
≠	領域Ｑ【Ｒｍ，ｍ≠２に対し
＜	実験しσ＜２で放電が　Ｅ／ｐ＜３０Ｖ／ｃｍ・Ｔｏｒｒ
＞	２温度プラズマ（Ｔｅ＞Ｔｇ）の
≦	を４８≦Ｅ／Ｐ０≦２０００Ｖｃｍ－１ｔｏｒｒ－１に対する立ち上がり時間≦８ｎｓ，
→	非品質合金→ＦｅｘＮｉ２３－ｘＢ６（Ｆｅ，Ｎｉ）固溶体
～	ＤＣアークに０．４　４０ｋＡ，１０－４　１０－７ｓの電流パルスを
＋	０．３３％Ｈｃｌ＋Ｈｅ（Ａｒ，Ｘｅ）中の
±	３６±０．５％ＭＭ含有ＭＭ－ＣＯ合金で

（５）数式処理

ＪＩＣＳＴ抄録では，第５表に示すように数式が多数日本語文中に埋め込まれた形で出現する．これを数式として認定することが非常に重要である．

文字列中に演算子として，＇＝＇，＇＜＇，＇＞＇，＇～＇，＇＋＇，＇≠＇等の記号列が見いだされるとその前後が数字列または英数字列からなる文字列に対しパターンマッチにより数式として判定しているしかし＇－＇はマイナスとしてだけでなくダッシュ，ハイフン等の機能として使われている場合が多く演算子として判定するのは危険である．数式処理は科学技術文献では重要であり，現在は形態素処理で上述のように行っているが構文的な処理を必要とする場合が多くある．

おわりに

３回目の実験が辞書の整備のため遅れており今回は報告できませんでした．今回未知語の登録と形態素解析能力の強化をはかったので処理の途中での失敗するケースの減少することが期待できる．

なお本実験に協力下された工業技術院筑波情報計算センターの渡辺定久氏，センター員の方々，オペレータ作業を担当したＳＡＫの落合衛氏およびＦＡＣＯＭ　ＳＥの方々に感謝致します．

参考文献

１）長尾　真，科学技術庁機械翻訳プロジェクトの概要，自然言語処理研究会資料３８－２，１９８３．

２）坂本義行　日本語形態素解析の基本設計，　同上

３）長尾　真，Ｍｕプロジェクトにおける日英翻訳結果の評価，　同上，４７－１１，１９８５．