自然言語処理 28－6
（1981. 11. 20）

テキサス大学における機械翻訳

首藤公昭（福岡大学・工）

1. はじめに

テキサス大学は、ALPACレポート以後も機械翻訳（MT）に情熱を持ちつづけてきた、アメリカでは数少い大学の一つである。今日、アメリカにおけるMTの商用システムとして、SYSTRAN, WEIDNER, LOGOS, ALPSなどが知られているが、大学関係に限れば、Brigham Young大学がMT研究から手を引いた現在、テキサス大学は、本格的にMT研究を行っている、アメリカで殆んど唯一の研究機関と云えそうである。本稿では、テキサス大学におけるMT研究の現状について紹介する。

2. プロジェクトの概要

テキサス大学言語学科に付属する研究機関であるLRC（Linguistics Research Center）は、1961年に設立され、以後、独語系言語に関する研究のかたわら、MT（主として独英MT）に欠要となる言語データの整理、及び翻訳方式や理論に関する検討が続けられたのち、1977年、米政府よりMTのための資金を得て、実験システムの作成及び実験が本格的に開始された。そして1980年、プロトタイプのシステムを完成して一応の区切りをつけたのち、現在、言語データ、プログラムの両面について拡張、改良作業が行われている。1982年中に次の段階の区切りをつけ、システムの総合的な評価を行うことになっている。プロジェクトのリーダーは、言語学科のW. P. Lehmann教授であり、言語的部分、プログラム部分の実質的責任者は、それぞれ、W. Scott Bennet と Jonathan Slocumである。その他、10名前後の言語あるいはコンピュータの専門家（院生も含む）がパート・タイムで働いている。現在のシステムは、INTERLISP versionである。大学内の5つのグループで共同利用できるDEC 2060（ALPANETも可）を数台の端末から使いながら研究を進めている。

3. 翻訳システムの概要

レキシコンや文法ルールなどの言語的部分は、将来にわたって常時改良され、拡張されるべきであるとする立場から、言語的部分をプログラム的部分から切離す様に配慮してあり、又、各部分でのモジュラリティも考慮されている。さらに、システム生成時に言語的部分を定義する際、言語の専門家に分り易い記述形式が定めてあり、コンパイルされてロー・レベル仕様に落される様になっている。目標とする所は、かなり広範な分野における、完全に自動化された多言語間翻訳としているが、当面は、データ通信関係の技術マニュアルを対象として独英翻訳の実用的システムを完成する事を目差している。システムの構成は、概略、図1に示すとおりである。

3.1 言語的部分

訳文を巧く作り出すために必要な情報を「特徴」（feature, 属性）、「特徴値」として整理している。（language-freeなものではない。）レキシコンがこれらによって記述されている例を図

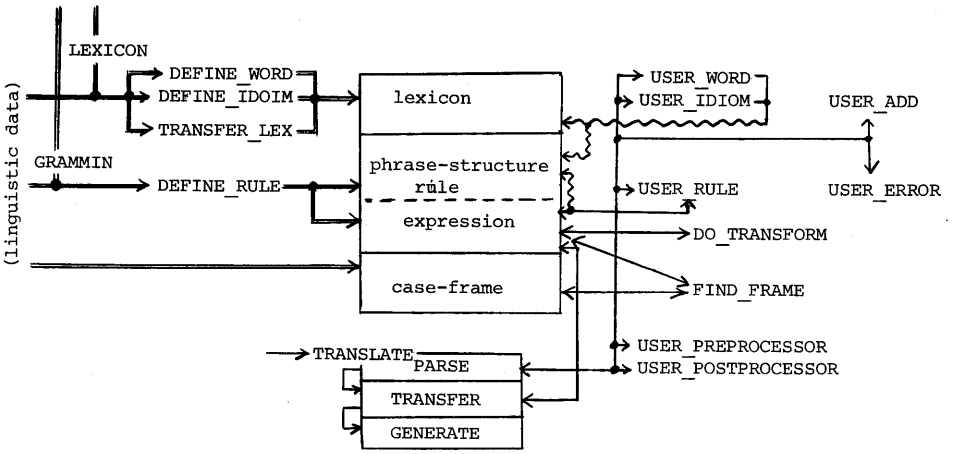

図 1. システムの構成

2に示す。基本型は OUTPUT、文法カテゴリー(CAT)は名詞語幹(NST)、異形態(ALO)としては output のみ、語尾変化(CL)は、単数形 S-01 又は複数形 P-S、capitalization (CP)は、通常行わない(LC)、始音(ON)は、母音(VO)、限定詞の必要性(DR)はどちらでもよい(RD NP)、性(SX)は unmarked (N)等々が与えられている。特徴(値)は、解析木を生成する際、上位ノードに対して、生成、伝播、消去のいづれかが行われる。又、特徴は、例えば、英語名詞には24種、独語動詞に24種用意されている。本システムは、句構造文法と格文法を基礎とし、木から木へのトランスファー方式を採用している。言語部は、大きく、レキシコン、句構造ルール、および格フレームから成る。

```
(OUTPUT              CAT (NST)
  ALO (output)
  CL  (S-01 P-S)
  CP  (LC)
  ON  (VO)
  DR  (RD NP)
  SX  (N)
  RC  (AGT)      (LOC)
  MC  (through)  (on)     $
  FC  (PP)
  PLC (WI)
  TAG (EWS)
```

図2. 特徴(値)による記述の例

3.1.1 レキシコン

レキシコンは形態素ベースであり、例えば、'gegangen' は、カテゴリー、GE-VB の 'ge', VST の 'gang' および V-FLEX の 'en' から一種の句構造ルール、VB ← GE-VB VST V-FLEX によって生成されたと見る。しかし、'Anschlussdatum' のように、形態素に分けると英訳 'line datum' となって正しい訳が得られない場合や、イディオムの場合は単一語としてエントリーすることもできる。また、対訳用辞書は、図3のような形式で与える。

```
[TRANSFER-LEX
( ON                    (AUF) PREP (PRCOM NIL))
( OUTPUT                (AUSGABE) NST)
( THE                   (DER) DET (KD DET))
( GO                    (GEHEN) VST (PX NIL) (PF FIN INF PAPL))
( MAGNETIC-TAPE         (MAGNETBAND) NST)
( AFTER                 (NACH) PREP (RO TMP))
( HOUR                  (STUNDE) NST)
]
```

図 3 トランスファー用レキシコンの例

3.1.2 文法ルール

各句構造ルールは、大きく4通りの表現によって強化されている。図4に独語の前置詞句を解析し、木のトラン

```
PP        PREP  NP
0         1     2
          --    (OPT KP * DEM)

TEST      (INT 1 GC 2 CA = X1)
          (OR (RET 1 WF)
              (AND (RET 2 BF) (INT 1 CN 2 BF = X2)))

CONSTR    (ADF 1 CAN PR)
          (ADF 1 RO)
          (ADX X1 CA)
          (CPX 2 CA NPCNJ)
          (AND (INT 2 KP REL) (ADD RELPRN))
          (AND (RET 2 NPCNJ) (ADD PPCNJ))
          (TAG ALL)

TRANSF    (SEF 1 CA GC)
          (OR (RET 1 PRCOM) (SEV 1 PRCOM NIL))
          (AND (INT 1 GC A) (INT 1 GC D) (SEV 1 GC D))
          (XFR)
          (CPY 2 MC)
```

図4 文法ルールと付記された表現の例

スファーを行うためのルールを示す。ルール強化部のすずcolumn test部は、句の構成要素それぞれに対して、より細かな制約を設けるもので、例では、名詞句NPが代名詞的特徴KPを持つならば、その値は定冠詞付き(DEM)であってはならない事を示す。次のtest部は、構成要素間のagreementをとるための条件で、例では、前置詞が名詞句に要求する文法上の格GCと名詞句の格CAとが一致すること(INTは特徴値集合の積演算)などの条件づけが行われている。constr部は、特徴とその値とを木の新しく作ったノードに付加するのが主な役割で、例では、前置詞の基本型CANの値を新ノードに付加し、それを特徴PRと名付けること、同様に、前置詞の役(role,深層格)ROを値と共に付加すること、さらに、test部で得た文法上の格の値を新ノードのCAの値とすること、等々が指定されている。最後のtransf部は、トランスファーフェーズで起動される部分であり、例では、PPノードのCA値を子のPREPノードのGC値とすること等々が指定されている。また、XFRによって子ノードに対するトランスファーを起動することが指定されている。

このルール強化部では、格フレームを参照して意味的なチェックを行ったり、トランスファー・フェーズで訳語動詞の性質に応じた語順を求めたりする関数FRMや木構造から木構造へ任意の深さで変換を行う関数、XFMなどが用意されていて、かなり柔軟に手続きを記述することができる。

3.1.3 格フレーム

roleの種類は、中心的(central)なもの9種、周辺的(peripheral)なもの30種程度が定めてあり、16種の格フレームが用意されていて、動詞のTT(transitivity type)特徴の値としていづれかが指定される。例えば、agentとlocativeを取るintransitive verb,'gehen'には、TT値としてI2ALが指定してあり、これは、図5に示すように定義された手続きである。図中、(1)の部分は、アナリシス・フェーズで、(2)の部分は、トランスファー・フェーズで働く。例えば図6の最上位ノードを生成するルール

```
(DEXPR I2AL (VC MD)
 (COND ((SYNTAX)
        (COND ((AND (ACTIVE)
                    (NON-COMMAND)
                    (FRAME N NP AGT)
                    (FRAME NIL NIL LOC))
               T)
              ((AND (ACTIVE) (COMMAND) (FRAME NIL NIL LOC)) T)))
       ((AND (ACTIVE) (NON-COMMAND) (PRES AGT LOC))
        (ROL-ORDER (AGT) (PRED) LOC))
       ((AND (ACTIVE) (COMMAND) (PRES LOC)) (ROL-ORDER (PRED) LOC))))
```

(1) (2)

図5. 格フレームの例

CLS ← PP RCL のtest部には関数FRMが書かれており、これから'gehen'のTT値、I2ALが起動される。PREDノードには、特徴として態(VC)は、能動(ACTIVE)、法(MD)は、直説法(INDICATIVE)が与えられているとすれば、図6の(1)の部分の第1サブフレームが成功し、主格(N)のNPノード('die Ausgabe')にRO値、動作主(AGT)が割り当てられ、'gehen'のlexical entryに指定されている、'auf'でマークされたPP句に当る'auf Magnetband'にRO値、位置(LOC)が割り当てられる。

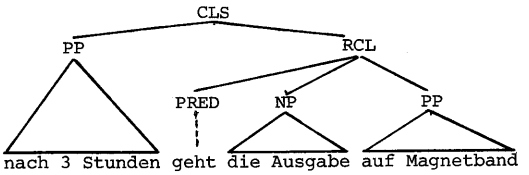

図6 格フレーム適用の例

3.2 手続き的部分

プログラムは、多くのモジュールから構成されており、トップレベルのTRANSLATE関数の起動によって翻訳過程が始まる。TRANSLATEは、基本的な3つのモジュールPARSE, TRANSFER, GENERATEを順に起動する。

3.2.1 PARSE

パージングは、Cocke-Kasami-Tounger のアルゴリズムを基本とし、これに"Top-Down Filter"と呼ぶ一種の予測機能(LINGOLの"oracle"と同じ)を加えることで、適用するルールの数を減らせる様になっている。(sub-grammer的な考えは採用していない。)ただし、この機能は、単語数の少い文に対しては、逆に能率を落すのでoptionalである(現在は単語数7以下の場合と解析が行き詰った場合はこの機能はoffされるようになっている。) PARSEは、文法ルールの適用に際して、直接ルール強化部を参照するのではなく、ユーザ定義可能な関数USER_RULEを起動することになる。レキシカル・アナリシスは、PARSEに起動されるUSER_WORDやUSER_IDIOMによって行われる。USER_WORDには、或る程度のリスペリング機能も組込まれている。レキシカル・アナリシスや、パージングに失敗すると、USER_ERRORというルーチンが起動され、ユーザーが必要とするapplication specificな手続きが定義できるようになっている。このように、パーザー本体をできるだけlanguage-specific, やapplication-specificな部分から切離す構成を採っている。

3.2.2 TRANSFER

トランスファー手続きの実質的部分は、各句構造ルールに付記されたtransf節の表現で示される。この部分には、関数XFRが書かれていて、これによって下位ノードに対するトランスファーが再帰的に起動され、終端ノードに至ってレキシカルトランスファーXLXが実行される。例えば、図6に関して述べた文法ルールのtransf部が図7で与えられるとすると、まずXFRによってそれぞれの子ノードに対するトランスファーを行ったのち、ORO によって図5のI2ALが起動され、今回は、(2)の第1サブフレームの部分が成功して訳文でのroleの順序AGT, PRED, LOCが定まる。

```
TRANSF  (XFR)
        (ORO)
```

図7

3.2.3 GENERATE

トランスファー・フェーズで得られた木をトラバースして終端ノードに至り、これを正しい形態に直して（レキシカル・トランスファーは基本型で行われている。）取り出し、left to right に並べてゆくものである。

3.2.4. その他の補助ルーチン群

レキシコンと文法ルールを定義するためには、言語の専門家に扱い易いようにハイレベルの記述を許す関数、LEXICON, GRAMMIN が用意され、よりロー・レベルの記述に対しては、DEFINE_WORDS, DEFINE_IDIOM, DEFINE_RULE などが使われる。又、頻繁に必要となる言語データの updating のためには、EDIT_WORDS, EDIT_IDIOMS, EXPUNGE_WORDS, EXPUNGE_IDIOMS, EDIT_RULES, EXPUNGE_RULES, EDIT_TXFORM, EXPUNGE_TXFORM など豊富な手続きが用意してある。

また、グローバル変数の初期化などの作業のためのユーティリティ、USER_PREPROCESSOR や、統計的な作業やパース結果にあいまいさの有る場合の選択作業などを行う USER_POSTPROCESSOR も定義することができる。

3.3 テキスト・マニピュレーション

現在対象としているテキストは、データ通信関係の技術マニュアルであり、原テキストとそっくり同じフォーマットの出力を得る様に考慮している。テキストの形式は図8に例示する様にかなりの図表を含んでいて、これらは翻訳対象とならない。そこで、まず、翻訳の準備として、翻訳対象となる単位を カッコ、[ ] でくくり、これらの単位の中に入り込んでいて無視されるべき部分をトグル・マーク、| でくくり、あるいは、数式などのように訳出する必要は無いが、文法上の単位とはなっている部分をマークしたりする必要がある。このためのプログラムが用意されている。

次に、トグルマークにはさまれた部分を省き、翻訳単位となる部分を抽出してファイルするスキャニング・デシジェーション・プログラムが用いられる。（図9, 10）

さらに、翻訳された結果を巧くならべて、原テキストに似たテキストを作る必要があり、このため、再構成プログラムを用いる。次に人手による修正作業をへて最終結果を得る。（図11, 12）

4. 実験結果の概要

1980年4月に行われた、対象とする技術マニュアル50頁に対する実験では、これらの頁をもとにして言語データが作られていた事もあって極めて良好であり、90%の文（翻訳の単位となる語列）が解析でき、83.7%の文は、post editionを必要としない程度に良い翻訳であった。その後、若干の改良の後、1980年11月に行われたブラインド・テストの結果を表1に示す。対象としたのは、上記とは異なる部分のテキスト751文であり 69.1%の文が解析可能、61.0%が post edition なしですむ良い訳出であった。誤った訳文を出した場合の原因は、約50%が文法ルールの未熟に起因するもの、約28%が文法ルールに付記されている preference factor の値が妥当でなく、誤った解釈を優先してしまったもの、約11%が文法ルールの不足によるもの、残りがレキシカル・アイテムのエラーによるものである。

なお、本システムは、入力ストリングが"文"として解析できない場合、最小数の句に分解した形で訳出を行うような fail-soft な機構を備えている（phrasal translation）。この機構が働くのは、レキシカル・エントリーの不備による場合が最も多

図8
原テキストの例

```
A.583.22        Satz-Buendel-Messung bei Fernbetrieb

        A.583.22

  Sollen alle Buendel    E1
  der VST gemessen werden?              N

           J
                          Soll die     E2
                          Ausgabe auf MB
                          erfolgen?              N

      MBG  01                   J

     Magnetband
     einhaengen

      DSS  02                                DSS  03

     Satz-Buendel-Messung             Satz-Buendel-Messung
     mit                              mit
     BA = BINAER                      BA = ASCII
     starten                          starten

                          Ende
I---------------I
I  A.583.22.E1  I  Sollen alle Buendel der VST gemessen werden?
I---------------I
     Wenn "Ja" muss die Ausgabe auf Magnetband erfolgen. Weiter
     mit S01.

     "Nein" bedeutet, dass nur ausgewaehlte Buendel der VST gemessen
     werden. Weiter mit E2.

I---------------I
I  A.583.22.E2  I  Soll die Ausgabe auf Magnetband erfolgen?
I---------------I

     Bei Messungen ausgewaehlter Buendel kann die Ausgabe auf
     Magnetband oder Drucker der TD-EWS erfolgen.
```

図9
準備

```
A.583.22        [Satz-Buendel-Messung bei Fernbetrieb]

        A.583.22

  [Sollen alle Buendel|    E1
  |der VST gemessen werden?]            N

           J
                          [Soll die|     E2
                          |Ausgabe auf MB
                          erfolgen?]             N

      MBG  01                   J

     [Magnetband
     einhaengen]

      DSS  02                                DSS  03

     [Satz-Buendel-Messung
     mit
     {BA = BINAER}
     starten]
                                      [Satz-Buendel-Messung
                                      mit
                                      {BA = ASCII}
                                      starten]

                          [Ende]
I---------------I
I  A.583.22.E1  I  [Sollen alle Buendel der VST gemessen werden?]
I---------------I
     [Wenn "Ja" muss die Ausgabe auf Magnetband erfolgen.] [Weiter
     mit S01.]

     ["Nein" bedeutet, dass nur ausgewaehlte Buendel der VST gemessen
     werden.] [Weiter mit E2.]

I---------------I
I  A.583.22.E2  I  [Soll die Ausgabe auf Magnetband erfolgen?]
I---------------I

     [Bei Messungen ausgewaehlter Buendel kann die Ausgabe auf
     Magnetband oder Drucker der TD-EWS erfolgen.]
```

かった様である。この実験でのテキストにおける1文当り平均語数は8.3、1文当り翻訳のCPUタイムは15.2秒（単語当り1.83秒）であった。

担当者の試算によれば、当初の5年間は、このシステムの場合、総計、年間31万ドル程度の費用が見込まれるが、人手による翻訳が1行当り1ドル位とすると、この費用は人手による毎分18語のペースでの翻訳に相当し、2～3人のpost-editorを考えに入れても、この程度の翻訳は十分、このシステムで行えると主張している。また、将来には人件費の上昇と計算機の低価格化が見込まれるので、MTは、益々、実際的な意味を持ってくると主張する。

4. むすび

本システムの主な特徴を列挙すれば、次のようになろう。

- 形態素ベース・レキシコン
- 属性－属性値システム
- 諸手続きで強化された句構造ルール
- 意味の格フレームによる取り扱い
- top-down的な予測を取り入れたCKYアルゴリズム
- 句構造ルールに強く結合されたトランスファー手続き
- モジュラリティー

```
(0763  Satz-Buendel-Messung bei Fernbetrieb)
(0764  Sollen alle Buendel der VST gemessen werden?)
(0765  Soll die Ausgabe auf MB erfolgen?)
(0766  Magnetband einhaengen)
(0767  Satz-Buendel-Messung mit {BA = BINAER} starten)
(0768  Satz-Buendel-Messung mit {BA = ASCII} starten)
(0769  Ende)
(0770  Sollen alle Buendel der VST gemessen werden?)
(0771  Wenn "Ja" muss die Ausgabe auf Magnetband erfolgen.)
(0772  Weiter mit S01.)
(0773  "Nein" bedeutet, dass nur ausgewaehlte Buendel der VST gemessen
       werden.)
(0774  Weiter mit E2.)
(0775  Soll die Ausgabe auf Magnetband erfolgen?)
(0776  Bei Messungen ausgewaehlter Buendel kann die Ausgabe auf Magnetband
       oder Drucker der TD-EWS erfolgen.)
```

図10 抽出された原文

```
(0763 peripheral circuit trunk group measurement in the case of
      remote operation)
(0764 should all trunk groups of the VST be measured ?)
(0765 should the output occur on MB ?)
(0766 hang up magnetic tape)
(0767 start peripheral circuit trunk group measurement with BA = BINAER)
(0768 start peripheral circuit trunk group measurement with BA = ASCII)
(0769 end)
(0770 should all trunk groups of the VST be measured ?)
(0771 if "Ja" the output on magnetic tape must occur .)
(0772 further with S01 .)
(0773 mean "Nein" , that only selected trunk groups of the VST are measured .)
(0774 further with E2 .)
(0775 should the output occur on magnetic tape ?)
```

図11 翻訳

```
A.583.22        Peripheral circuit trunk group measurement in the case of
                remote operation


          A.583.22

   Should all trunk groups  E1
   of the VST be measured?          N

                J
                                Should    E2
                                the output
                                occur on MB?          N

             MBG  01                 J

          Hang up
          magnetic tape


             DSS  02                                  DSS  03

          Start peripheral                         Start peripheral
          circuit trunk group                      circuit trunk group
          measurement with                         measurement with
          BA = BINAER                              BA = ASCII

                                  End

I---------------I
I  A.583.22.E1  I  Should all trunk groups of the VST be measured?
I---------------I
      If "Ja" the output must occur on magnetic tape.
      Continue with S01.

      "Nein" means that only selected trunk groups of the VST are
      measured. Continue with E2.

I---------------I
I  A.583.22.E2  I  Should the output occur on magnetic tape?
I---------------I

      The output may occur on magnetic tape or on the printer of the
      TD-EWS in the case of measurements of selected trunk groups.
```

図12 最終結果

- 豊富なユーティリティ
- 文解析不能の際、句解析に切りかえたり、綴りの誤りを正す fail-soft 機構
- テキスト・マニピュレーションの自動化

研究の基本的な姿勢は、例えば、知識ベースのわく組みをとり入れるといった冒険的な方向を避け、従来、比較的良く知られている言語学上、コンピュータ工学上の結果をいかに巧く組合せて、できるだけ実用的なMTを実現するかを重視する立場をとっている。当面はヨーロッパ系の言語しか対象としておらず、日本語のような大幅に異なる言語に彼らのシステムがどこまで有効かについては疑問が残っている。彼らのシステムでは、中間表現(intermediate expression)の考えが薄く、かなり表層に近いトリーから直接、ターゲットの表層トリーに変換する方式をとっており、真の多言語間翻訳のわく組みとしては、この点が問題であろう。概念間のリレーションに基づいたトランスファーがぜひとも望まれる所である。又、semantic featureも未だ完全に整備されていない様に見うけられる。将来、この種の意味の取り扱いが整備され十分に機能した際、どの程度のperformanceが得られるかは、興味深い。

パージングの方式としては、非文法的文が入力される事も多い実際の現場では、

本システムのようにbottom-up方式を採用するのが妥当と思われるが、将来の文法ルール等の大幅な増補を考えると、やはり、実行段階でfilterによって適用ルールを減らすだけでなく、sub-grammar的な考えを採用する事が必要ではなかろうか。(現在のルール数は300程度であるが、将来2,000程度までふくらむであろうと予測されている。)

以上のような、いくつかの基本的な問題点は感じられるが、ヨーロッパ系の言語間で、分野を限定すればある程度の成功をおさめるのではないかと思われる。第2世代の末尾に位置し、第3世代への橋渡しの役を担うシステムの一つとして、今後のなり行きが大いに注目される所である。なお、本研究プロジェクトが言語学者によって創始され、主催されている事もComputational Linguisticsの在り方として参考になる点であろう。

Original Document:	751 sentences (6216 words)	
	32 pages	
	23.5 sent/page	
Raw results		
Analyzed	519 sentences	(69.1%)
Good transl.	458 sentences	(61.0%)
Bad transl.	61 sentences	(8.1%)
Not analyzed	232 sentences	(30.9%)
Phrasal paren.	23 sentences	(3.1%)
Phrasal transl.	153 sentences	(20.4%)
Too many phrases	18 sentences	(2.4%)
Too much space	38 sentences	(5.1%)
Full analysis		
Major source errors:	8 sentences	
Computation base:	743 sentences	
Analyzed	519 sentences	(69.9%)
Good transl.	458 sentences	(61.6%)
Choices ind.	9 sentences	(1.2%)
Bad transl.	52 sentences	(7.2%)
Not analyzed	224 sentences	(30.1%)
Missing lex. entries	59 sentences	(7.9%)
Perf. phrasals	5 sentences	(0.7%)
Good phrasals	4 sentences	(0.5%)
Lesser phrasals	101 sentences	(13.6%)
No transl.	55 sentences	(7.4%)

表1 実験結果

文献

- W.P. Lehmann, W.S. Bennett, J. Slocum, et al: "THE METAL SYSTEM", RADC-TR-80-374, Vol I, Vol II, 1981
- LRC: monthly reports, 1977~
- 長尾: "ヨーロッパにおける機械翻訳の現状", 情処研資, CL 20-1, 1979
- 長尾: "機械翻訳", 情報処理, 20, 10, 1979
- V.R. Pratt: "LINGOL - A PROGRESS REPORT", IJCAI, 1975