英文補完システムの補完能力評価

唐沢 博　小川 均　田村 進一
（大阪大学 基礎工学部）

1. まえがき

従来より幾多の自然語処理システムが構築され報告されてきたが、そのシステムが採用している手法の適用範囲を見定めるのに必要な評価は、ほとんど例がない。このため自然語処理手法の効果がいまひとつ明らかに理解できず、これが自然語処理研究全体に渡る不透明さとして受け取られている傾向がある。こういった傾向は、自然語理解の研究において特に著しい。自然語自身が閉じた系ではないという性質とも相俟って、システムの能力評価は一般に難しい。〝理解〟のプロセスが関与するようなシステムの評価は、特に困難である。

筆者等は、人間が頭に浮かぶ単語，句，文などの集合からなる情報（以後、これを不完全テキストと呼ぶ）を入力すると、推論の及ぶ範囲内で必要な欠落情報を補って文法的に完全な英文テキストを出力する、英文テキスト補完システムを既に作成し報告した。[1] そして同システムの補完能力を評価するにおよんで、適切な一般的手法がないことを知り、そのような方法論の提案も含めて補完能力の評価法を考案し、同システムに適用した。その結果、少なくとも補完システムにとって、その評価法が有効であることがわかった。本論では、同評価手法および評価結果、さらにシステムの tuning に適用する方法について報告する。

2. 補完システムの概要

先ず補完システムの全体的な構成と処理プロセス、および補完処理について概要を説明する。

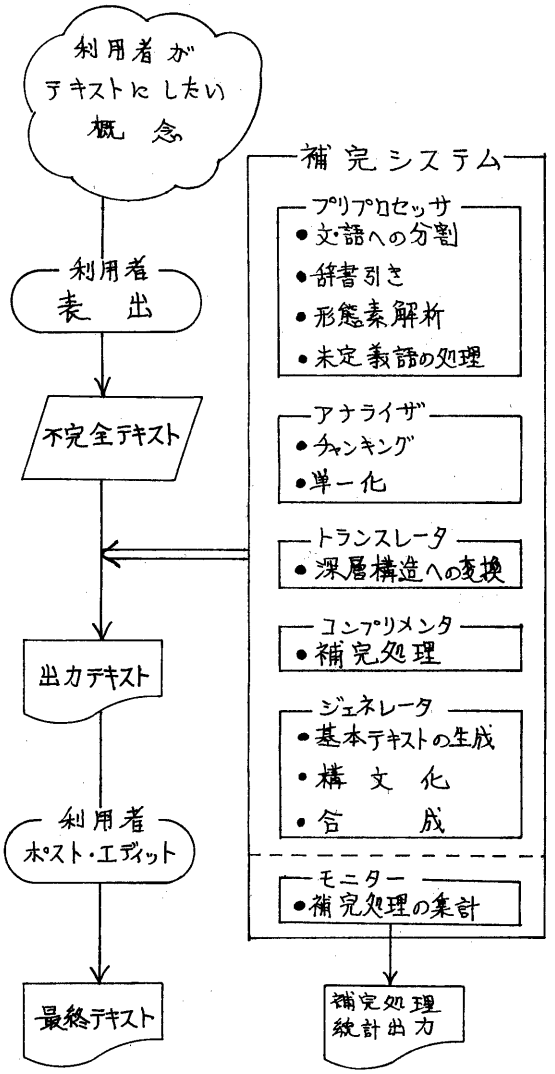

図1　システム構成

2.1 全体的な構成

図1に示したように、補完システムは、入力された不完全テキストに対し前処理を施してから解析を行い深層構造へ変換する。この深層構造に対して補完処理が実行され、その結果を表層

構造としての英文テキストに変換して出力する。利用者は、この出力テキストにポスト・エディットを加え、目的とする最終テキストを得る。補完システムの目的は、利用者の英文テキスト作成の支援にある。図2に各プロセスにおける処理結果の例を示す。

2.2 補完処理

補完は「補足」と「付加」とからなる。これらは、必須要素であるにもかかわらず欠落しているような情報を、統語的知識や一般的知識を用いて補足する手続きと、存在すれば利用者にとって有益であろうと予測される情報の付加の手続きとである。補完はさらにその段階において「第1次補完」と「第2次補完」とに大別される。前者は統語レベルの浅い補完であり、統語構造上必要とされる項目に対し、暫定的なマーキングを行う。後者は種々の知識を用いた深いレベルの補完であり、これを実現するために4種類の補完型が存在する。補完型は、文脈型・連想型・推論型・デフォルト型があり、プロダクション・システムに類似した機構を持つ。表1に各補完型の特性を示す。

表1 各補完型の特性

文脈型	冗長な繰り返しをさける目的で省略された情報を前後の文から復元する。
連想型	入力情報中の語をキーとして連想される情報を辞書中から取り出して省略を補う、もしくは冗長な情報として付加する。
推論型	一般的な知識（世界知識や常識的知識）を用いて、意味的に省略されている情報を復元する。
デフォルト型	上記3種類の補完型でも文要素が復元できなかった場合に、常識的な範囲で文要素を補う。

```
I GOT UP AT 7, AND SCHOOL AT 8
MATHEMATICS AND ENGLISH
AT HOME 4.   SUPPER

a) input text

1.  I GET-UP AT 7 , AND SCHOOL AT 8
2.  MATHEMATICS AND ENGLISH
3.  AT HOME 4 .
4.  LAST-MEAL-OF-THE-DAY

b) preprocessed text

1.  I GET-UP AT 7 : T10
2.  SCHOOL AT 8 : , RPA
3.  MATHEMATICS-AND-ENGLISH
4.  AT HOME 4 .
5.  LAST-MEAL-OF-THE-DAY

c) analized text

1.  S:I  V:GET-UP  TENSE:T10  TIME(1):AT-7
2.  RELATION:RPA  PLACE(1):AT-SCHOOL
TIME(1):AT-8
3.  O:MATHEMATICS-AND-ENGLISH
4.  PLACE(1):AT-HOME  TIME(1):AT-4-O'CLOCK
5.  O:LAST-MEAL-OF-THE-DAY

d) pre-complementing deep structure

1.  S:I  V:GET-UP  TENSE:T10  PLACE(1):AT-HOME
TIME(1):AT-7  STARTING-POINT:FROM-THE-BED
2.  S:I  V:GO  TENSE:T10  RELATION:RPA
PLACE(1):AT-SCHOOL  TIME(1):AT-8
STARTING-POINT:HOME  TERMINAL:SCHOOL
WAY:BY-SOME-WAY
3.  S:I  V:HAVE-EDUCATIONS-OF
O:MATHEMATICS-AND-ENGLISH  TENSE:T10
TIME(2):WHILE-I-BE-AT-SCHOOL
COMMENT(1):MOD/DIFFICULT  COMMENT(2):GTIRED
4.  S:I  V:GO-BACK  TENSE:T10
PLACE(1):AT-HOME  TIME(1):AT-4-O'CLOCK
STARTING-POINT:SCHOOL  TERNINAL:HOME
WAY:BY-SOME-WAY
5.  S:I  V:HAVE  O:LAST-MEAL-OF-THE-DAY
TENSE:T10  PLACE(1):AT-TABLE
TIME(1):IN-THE-EVENING  COMMENT(2):FHAPPY

e) complemented deep structure

I GOT UP FROM THE BED AT 7, AND I WENT BY SOME
WAY TO SCHOOL FROM HOME AT 8.  I HAD
EDUCATIONS OF MATHEMATICS AND ENGLISH THAT
WERE DIFFICULT WHILE I WAS AT SCHOOL, AND I
GOT TIRED.  I WENT BACK BY SOME WAY TO HOME
FROM SCHOOL AT 4 O'CLOCK.  I HAD LAST MEAL OF
THE DAY AT TABLE IN THE EVENING, AND I FELT
HAPPY.

f) output text
```

図2 処理例

3. 評価Ⅰ——各補完型の寄与率に関する評価

この評価は、補完体系の特性を明らかにすると共に個々の利用者に補完システムをtuningするための基礎を与える。i)第2次補完を実行する4種類の補完型が、どんな割合で寄与しているのかという点、およびii)質の良い出力テキストを生成する時、どの補完型が効いているのかを知る目的をもつ。[2)] 評価は、第1次補完および文脈型，連想型，推論型，デフォルト型を対象とした。

3.1 評価Ⅰの実験系

図3は、評価Ⅰのための実験系である。試行者が表出した英語の断片情報を不完全テキストとしてシステムに入力する。システムは補完を行った後、テキストを出力する。もとの試行者はこの出力テキストを読み、自分の意図していた内容を表わしているかどうかで、質の良いテキスト群と良くないテキスト群とに大別する。システムはまた、各補完型の起動回数と比率に関する統計情報も各出力テキストとともに出力する。

3.2 評価Ⅰの結果

図4の(a)に質の良いテキスト群の結果を、(b)に質の良くないテキスト群の結果を示す。これは、個々のテキストとともに出力される統計情報を各群毎に集計したもので、各補完型の数値はすべての補完型の平均起動回数に対する、その補完型の平均起動回数の百分率を示している。(a)と(b)を比較することにより、推論型や連想型が多く寄与する程テキストの質が向上する、換言すれば人間の意を良く汲み取ることが

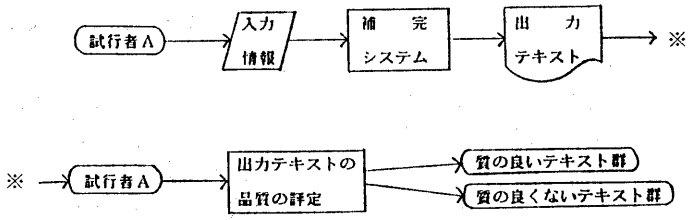

図3 各補完型の寄与率を評価する実験系

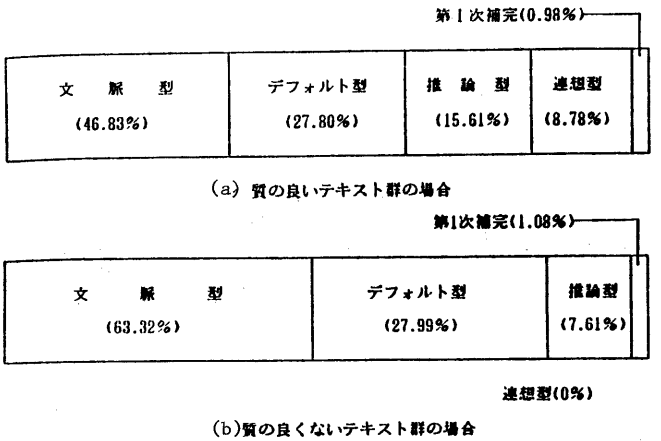

図4 評価Ⅰの結果

わかる。

この結果から省略の性質を考えてみると、次のことが言える。すなわち、一般的知識を用いて推論可能な物事や容易に連想されるような語、句、文の省略は比較的主観的に行われ、それらがうまく復元できた時の効果は大きい。

従って、コミュニケーションを円滑に実現させるために必要な範囲の一般的知識の集合と連想事項の集合を見極めることが大切である。しかしこの集合は個々の利用者に依存するし、また同一の利用者でも、その話題に依存する。こういった依存性を考慮した上で上記の集合を決めるために、後述するtuning 技法を用いる。

4. 評価II ——人間の補完能力との比較評価

システムの補完能力が、人間の補完能力と比較してどの程度なのかを知る目安を得るための評価である。[2)]

4.1 評価IIの実験系

試行者が表出した英語の不完全テキストを、システムと英語圏の人間（以後、nativeと呼ぶ）の両方に与える。システムが生成した出力テキストと、nativeが作成したテキストの両方を、もとの試行者が自分の意図に合う内容にするべくポスト・エディットする。ポスト・エディット量に関して、システムの出力テキストに対するポスト・エディット量をα、nativeの作成テキストへのそれをβとおく。このとき、

$$R_H = \frac{1+\ln(\beta+1)}{1+\ln(\alpha+1)} \qquad (1式)$$

を補完比と呼び、人間との補完能力もしくは補完システム間の補完能力の比較の目安とする。

$$\alpha \geq 0,\ \beta \geq 0$$

であるので、

$$\alpha+1>0,\ \beta+1>0$$

として対数式の真数が0になるのを防いでいる。さらに、

$$1+\ln(\alpha+1) \geq 1$$

だから、1式の分母は0にならない。分子の形は、分母にあわせた。ポスト・エディット量の対数をとった理由は、人間の主観や感覚が物理量の対数に比例するとするWeber/Fechner の法則[3)]に従ったことによる。

R_H は次の様な意味を持つ。

i) $R_H < 1$ ...システムの補完能力が人間よりも劣る。$0 < R_H$ なので、0に近い程、能力が低い。

ii) $R_H = 1$ ...システムと人間の補完能力は同等である。

iii) $1 < R_H$ ...人間よりもシステムの補完能力が優れている。$R_H < \infty$ であるので、そのレンジで考慮する。

一方、ポスト・エディット量は次の規則に従って算定した。

1)補完システムの解析部は、意味のかたまり（チャンク）として断片情報

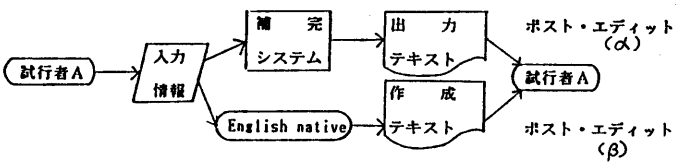

図5　システムの総合的な補完能力を評価する実験系

をとらえる。従って、文，句，単語は、それぞれがエディットの単位とされた場合、いずれもカウントを1とする。

2)挿入は1，削除は0，置換は削除と挿入の組合せと考えて1とする。

また、入力情報の多少が出力側のポスト・エディット量に影響を与えないように、補完システムとnativeとの2系統を通すことによって打ち消し合うよう考慮した。図5に実験系を示す。

4.2 評価Ⅱの結果

40例のテキストについて補完比を求めた結果を表2に示す。最大値1.0は$\alpha=\beta=0$のテキスト例であった。そのテキストへのポスト・エディットは、冗長な情報の削除のみだった。最小値を示したテキストは、その内容がほとんど意味をなさなかった。

4.3 補正評価値

補完比は、それ自身の解釈は$R_H=1$の場合以外、単なる目安程度の役割しか果さない。そこで、補完比に補正を加えて単独値でも有益であるように拡張した。

すなわち、nativeの作成したテキストを正答（100点）として、それに対するシステムの出力テキストに0点～100点の範囲で主観採点値Pを与える。これと補完比との間には相関があることが予想される。Weber/Fechnerの法則に基づき、

$$\ln\left(100\times\frac{\beta+1}{\alpha+1}\right) \qquad (2式)$$

とPとの相関を調べたところ有意であったので、

$$P=k\cdot\ln\left(100\times\frac{\beta+1}{\alpha+1}\right)+b \qquad (3式)$$

とおいて回帰的にk,bを次式を得た。

$$P=57.83\ln\left(100\times\frac{\beta+1}{\alpha+1}\right)-167$$

$$\fallingdotseq 57.83\ln\frac{\beta+1}{\alpha+1}+99.32$$

但し、$(\beta+1)/(\alpha+1)<0.18$の場合、

表2　評価Ⅱの結果

	補完比	補正評価値(点)
最大値	1.0	99.3
平均値	0.58	35.8
最小値	0.32	0

$P<0$となってしまうので結局、次式を実際に用いる。

$$\begin{cases} P=57.83\ln\dfrac{\beta+1}{\alpha+1}+99.32 & \left(0.18<\dfrac{\beta+1}{\alpha+1}\right) \\ P=0 & \left(0<\dfrac{\beta+1}{\alpha+1}<0.18\right) \end{cases} \qquad (4式)$$

4式によって求められるPの値は、システムの出力テキストを仮りに採点した場合、何点ぐらいの得点をとるような能力を持つかの指標として用いることができる。このPを評価Ⅱの補正評価値と呼ぶことにする。表2には、補正評価値が補完比とともに示されている。

5. 補完システムのtuningへの適用

評価法のⅠおよびⅡを用いることによって、より質の高い出力テキストが得られるよう補完システムをtuningすることが可能である。評価Ⅰの手法を用いて，システムの辞書内容、知識ベースの内容を、より適切なものに更新する。そして評価Ⅱの手法によって総合的な改善度を確認する。

5.1 評価Ⅰの手法の適用

各補完型は、それぞれ以下のような知識もしくは情報と関連している。

1)オ1次補完…文法知識

2)オ2次補完

2-1) 文脈型…統語構造上での省略の行われ方に関する知識.

2-2) 連想型…一般的に連想さ

れるような語，句，文を示す辞書項目．

2-3) 推論型…推論時に起動される規則形式の一般的な知識．

2-4) デフォルト型…常識的に省略されていると考えられる文要素を補うための知識．

3章で論じたように、連想型，推論型の補完が寄与する程、出力テキストの質は良かった。従って、より質の良い出力テキストを得るためには特に，

a)辞書項目内で各単語の属性情報として存在する連想情報を、より適切な内容にする．

b)推論に使われる一般的な知識の内容を検討する。

図6に示したように、評価Ⅰの結果をシステム内（辞書内容，知識ベースの内容）にフィードバックして利用者の使用環境に適合させることにより、tuningを実現させる。これは、知識獲得という観点からは、教師付き学習である。

全出力テキスト数をN_A，質の良いテキスト群の出力テキスト数をN_G，質の良くないテキスト群のそれをN_Bとすると、

$$N_A = N_G + N_B$$

の関係にあるので、基本的には、

$$N_B = 0$$

すなわち、

$$N_A = N_G \qquad (5式)$$

となるまでフィードバック・ループを繰り返せばよい。

5.2 評価Ⅱの手法の適用

前節の系のもとでtuningされた補完システムが、システム全体として総合的な補完能力がどれ程改善されたかを知るために評価Ⅱの手法の適用を行う。

今、ある時点もにおいて評価Ⅰの手法の適用によるtuningが終了した時の補完比をR_H^G，しばらくその状態でシステムを利用した後、再びtuningの

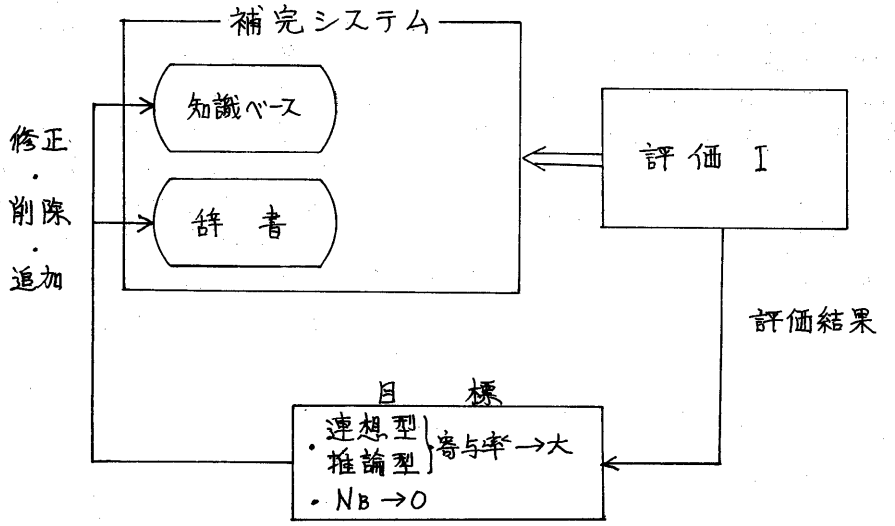

図6　評価Ⅰの手法の適用によるtuning系

必要が生じて、評価Ⅰの手法により再tuningを終えた時点t+1の補完比をR_H^{t+1}と表現すれば、

$$R_H^{t+1} > R_H^{t} \quad (6式)$$

という関係が成立する限り、システムの総合的な補完能力は改善されていると考えてよい。あるいは、

$$\tilde{R}_H = 1 \quad (7式)$$

または、

$$\tilde{P} = 100 \quad (点)$$

であれば、システムは、その状態で当面は実用上の運用に耐えるとみなしてよいであろう。

5.3 補完操作トレース情報の利用

今まで述べてきたtuningによっても5式，6式，7式等が達成できない場合もある。これは、その原因が補完処理の制御に及ぶもので、主として各補完型の起動プロセスの適否に関係している。（別な要因は「考察」の章で述べる。）補完処理を実行するモジュールであるコンプリメンタは、補完操作に関するトレース情報を出力するモードを有する。この情報は、いかにして補完操作を行ったかの説明機能ともみなせる。補完に関与した全知識規則が、起動された順番に、どこの情報を参照して、どのような情報を生み出したかをも併せて出力される。tuning技法のback-up的な手法は、この情報を活用する。すなわち、コンプリメンタにbugがあるとみなし、tuningをdebugのプロセスに対応させる。この最もコスト高のtuningは、補完時に使用される各知識の適用方略に関する知識（メタ知識）の最適化であるといえる。

尚、トレース情報は、評価Ⅰの手法の適用によるtuningにおいても、変更すべき知識単位がどれであるかを知るためにも利用される。

6. 考察

評価Ⅰの結果は40例のテキストを用いて、一度もtuningしていないシステムに対して得たものである。この場合、

$$N_G / N_B \simeq 0.8$$

であったから、システムの適応状態はあまり良いとは言えない。

また、図3の系では入力情報の量の影響が出力側へ出てしまう。入力情報の断片数は5～10個に制限したが、この影響は確かに出ており、図5の系で初めて中和されるのが確認された。この点に関しては、検討の余地がある。

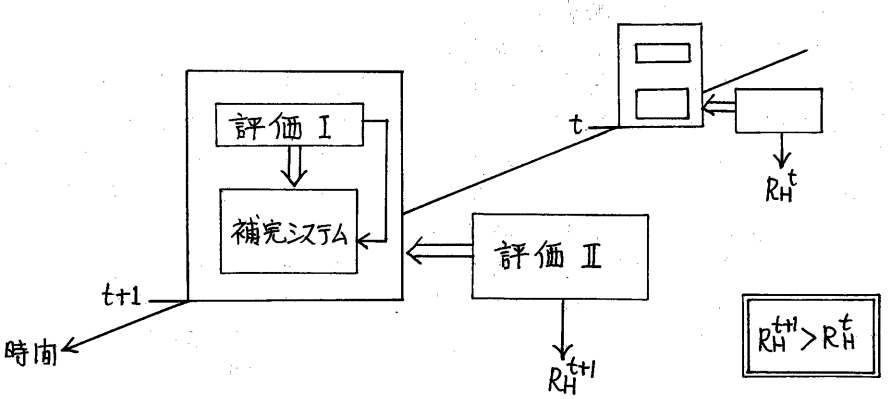

図7 評価Ⅱの手法の適用によるtuning系

しかしながら、内容的に質の良いテキストを出力する際に、強力に寄与しているのが連想型，推論型であることが明らかとなった。この事実は、省略現象そのものの性質、および補完システムの補完体系の性質を物語っている。

評価Ⅱの手法はかなり有効であり、この種のシステムの能力を数量的かつ1元的に評価する1つの方法を示せたように思われる。但し、ポスト・エディット量の算定の仕方が粗すぎるのではないかという指摘があり、例えば、文，句，単語のレベルでテキスト全体への各々の貢献度が違うのならば、それを考慮したグレードを導入する必要があろう。

5.3節でふれた様に、tuningが効果を示さない別の大きな要因として、現在の補完体系が、はたして省略現象全体をカバーしているかという点がある。例えば現行の体系では、談話理解のプロセスが浅い。文脈型は、統語レベルに近い文脈情報の処理が中心であり、推論型もコンテクスト全体との関係において機能している規則はほとんどない。従って、ストーリのスキーマを処理するような補完型の導入は、補完システムの能力を大きく向上させることが予想される。これには、スクリプトやMOPsといった知識表現形式が用いられるだろう。もう1つ重要なことは、定形文書のように表層レベルにおける紋切り型表現の存在である。出力テキストの形式を、いずれかの定形表現に落とすということは、テキストの質の問題も含めて、システムの実用性をさらに高める意味で重要である。この点については、京大工学部の辻井潤一助教授より御指摘があった。

7. まとめ

利用者が頭の中に描いた内容と補完システムが物理的に生成したテキストの内容との比較評価を行うために、心理学的な手法を導入した。システムの出力テキストの内容は、入力情報に対する1つの解釈であり、この解釈が利用者のあらかじめ意図していた内容にどれだけ近いかを評価するのが、本論における評価の目的である。この評価により、補完体系の性質を明らかにしたと同時に、同評価手法の応用として、質の良い内容のテキストをより多く生成させるためのtuning技法へ発展させた。

最後に、心理学的見地から種々の有益な助言を頂いた京大文学部心理学教室の乾敏郎助手，および評価実験用に補完システムを拡張してくれた大学院生の堂坂浩二君に感謝します。また、この研究の一部は文部省科学研究費による。

参考文献

[1] 松永，小川，田中，"マイクロ・コンピュータ上での補完的英文生成システムの実現"，情報処理「知識工学と人工知能」研究会資料，28-2，1982

[2] 唐沢，田村，松永，"英文補完生成システムの補完能力評価"，情報処理学会第27回全国大会予稿集，4D-3，pp.1139-1140，1983

[3] Rumelhart, D.E., "HUMAN INFORMATION PROCESSING", John Wiley & Sons, Inc., 1977