自然言語処理　138－5
（2000．7．18）

コーパスからの省略補完ルール獲得環境

中岩浩巳
ＮＴＴコミュニケーション科学基礎研究所
〒619-0237 京都府相楽郡精華町光台 2-4
nakaiwa@cslab.kecl.ntt.co.jp

あらまし　本稿では，単言語コーパス及び対訳コーパスから効率的に省略補完ルールを獲得するソフトウェア環境を提案する．本環境では，特定のコーパス向けに省略補完ルールを作成する過程を考慮にいれて，一般的に入手容易な日本語単言語コーパスから省略箇所とその補完要素の情報をタグ付けした省略補完タグ付き n 日本語コーパスを効率的に作成し，その結果をもとに省略補完ルールを効率的に作成する．また，機械翻訳システムでの適用を想定した状況では入手が比較的容易な日英対訳コーパスから，日本語文中の省略箇所とその英語文中の補完要素を自動抽出し，その結果を元に自動的にルールを獲得する機能も備えている．本環境は，日本語解析系として日英機械翻訳システム ALT-J/E を活用して実装している．

キーワード　省略，ゼロ代名詞，補完，照応解析，コーパス，機械学習

An Environment for Extracting Resolution Rules of Zero Pronouns from Corpora

Hiromi Nakaiwa
NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Souraku-gun, Kyoto 619-0237 Japan
nakaiwa@cslab.kecl.ntt.co.jp

Abstract　This paper proposes a practical integrated environment for extracting rules for the anaphora resolution of zero pronouns from monolingual and/or bilingual corpora. This method takes into account the practical situation for making resolution rules of zero pronouns in specific domain texts; the types of usable corpora (monolingual and/or bilingual) for examining the extraction of resolution rules have been changed depending on the type of NLP system using extracted resolution rules. The extraction processes of resolution rules in the environment are classified into five component tasks: (1) Zero Pronoun Identification, (2) Antecedent Annotation, (3) Rejection of Sentences Unsuitable for Rule Extraction, (4) Rule Extraction, and (5) Extracted Rule Application and Modification. An automatic process and/or a manual process with a user friendly human interface can be used to achieve each component task. This environment was implemented in the Japanese-to-English machine translation system, ALT-J/E, for Japanese zero pronoun resolution.

key words　ellipsis, zero pronouns, supplementation, anaphora resolution, corpora, machine learning

1. はじめに

自然言語では通常，相手（読み手もしくは聞き手）に容易に判断できる要素は，文章上表現しない場合が多い．この現象は，機械翻訳や文書要約等の自然言語処理において，大きな問題となる．例えば，機械翻訳では，原言語では陽に示されていない要素が，目的言語で必須要素になる場合，陽に示されていない要素の同定が必要となる．特に日英機械翻訳システムにおいては，日本語の格要素が省略される傾向が強いのに対し，英語では訳出上必須要素となるため，この省略された格要素（ゼロ代名詞と呼ばれる）の照応解析技術は重要となる．

日本語省略格要素の補完処理に関しては，従来から様々な手法が提案されてきているが[1][2][3][4]，翻訳対象分野を限定しない機械翻訳システムに応用することを考えると，解析精度の点や対象とする言語現象が限られる点，また，必要となる知識量が膨大となる点で問題があり，実現は困難である．これらの問題に対しては，照応解析条件として，用言の意味属性[5][6]，様相表現，接続表現を用い，これらを表現の持つ意味に応じて分類し，その代表的属性値に応じて照応要素を決定することによりこれらの問題を考慮にいれた，機械翻訳に適した省略補完手法が提案されている[7][8][9]．

しかし，これら従来から提案されている手法では，基本的に人間が省略補完のためのルールを作成する必要がある．よって，網羅的な省略補完ルールを作成するためには，かなりの専門知識と労力が必要となる．さらに，解析対象となる分野に応じて，異なった要素を補完する必要がある省略箇所が存在するので，分野に依存した省略補完ルールを作成する必要がでてくる．しかし，分野毎にこのルールを作成することは，その労力や時間を考慮すると，実際的には実現不可能である．よって，この省略格要素の補完ルールを効率的に獲得する手法の実現が望まれている．

自然言語処理システムにおける解析規則の効果的な獲得のためには，従来から，解析対象言語コーパスから省略補完ルールを獲得する手法[10][11]，対訳コーパスから獲得する手法[12][13]，省略箇所に対する補完要素がタグ付けされたコーパスから獲得する手法[14][15]が提案されている．

これら手法のうち，解析対象言語のコーパスのみを使用する場合，その言語ではほぼ常にゼロ化される要素を補完するための規則を抽出することは困難である．

これに対し，日本語と英語のように言語族が異なる対訳コーパスでは，省略される傾向が異なるため，ある言語の文では省略されている要素が，その文と対訳関係にある別の言語の文では明記される場合が多々あり，その利用が有効である．しかし，対訳コーパスはその入手が比較的困難であるという問題がある．

補完要素がタグ付けされたコーパスから獲得する手法では，タグ付けされた情報を手掛かりに有効なルールを効率的に獲得することが可能である．しかし，現時点ではこのようなタグ付けされたコーパスはほとんど無く，タグ付け書式の標準化にむけた取り組みが行われている段階である[16]．よって，実際の場面では，省略補完ルールを作成するアナリストが省略箇所と補完要素の情報を人手作業でコーパスにタグ付けするのが現実的な状況である．以上のことから，効率的に省略補完ルールを獲得するためには，テキスト中の省略箇所に対する補完要素の情報を効率的にタグ付けするツールが必要となる[17]．

特定分野テキスト向きの省略補完ルールを作成する際には，通常は，省略補完タグの付与されていない単言語コーパスを分析して作成する．よって，アナリストはまずコーパス中の省略箇所に対する補完要素のタグを効率的に付与する必要がある．しかし，機械翻訳向けのルールを作成する場合には，その特定分野の対訳コーパス（例えば，前の版の原稿の翻訳結果や翻訳メモリーシステム中の対訳コーパスデータ等）が活用できる場合が多い．このような状況では，対訳コーパスから省略補完ルールを自動獲得する手法[14][15]が活用できる．しかし，この自動獲得手法では，最適のルールが獲得できる保証はない．よって，高い補完精度が要求される状況では，自動獲得した翻訳ルールを人間が確認する必要がある．さらに，ルール獲得の際に必要となる各種人手操作は，単言語コーパスからでも対訳コーパスからでも，効率的に行なえる設計となっていなければならない．

以上のような省略補完ルールを獲得する実際の状況を考慮に入れて，本稿では，単言語および対訳コーパスから効率的に省略補完ルールを獲得する実用的な統合ツールを提案する．

2. 省略補完ルール獲得のサブタスク

コーパスから省略補完ルールを獲得する際に必要となるサブタスクは大きく次の5項目に分類できる：(1)省略箇所認定，(2)補完要素認定，(3)ルール獲得不適文の排除，(4)ルール抽出，(5) 獲得ルールの適用と修正．

2. 1 省略箇所認定

省略箇所の認定では，獲得したルールを活用する自然言語処理システムにおいて補完処理が必要となる省略箇所を認定する．例えば，日本語のような語順が比較的自由な言語では，必須となる格要素を特定する明示的な手掛かりがない．よって，このような言語では，コーパス中で補完が必要な省略箇所を認定する処理が

省略補完ルール獲得において重要である．さらに，補完が必要な省略箇所は自然言語処理システムのタイプに応じて異なってくる．例えば，機械翻訳では，目的言語で明示して訳す必要のある省略箇所だけに補完すればよい．例えば次の日本語文(1)では，主語（ガ格）が日本語では明示されていないが，英訳すると，"Zoos raise lions"という表現で翻訳することができるので，機械翻訳ではこの主語は補完が必須な要素とはならない．

(1) (ϕ-が) 動物園でライオンを飼う．
"Zoos raise lions."

以上のように省略箇所の認定においては自然言語処理システムの解析結果を考慮にいれなければならない．

単言語コーパス中の省略箇所認定は，自然言語処理システムの解析結果だけで行える．これに対し，対訳コーパスでは，省略箇所に対する翻訳結果も補完が必要な省略要素を認定する手掛かりとして活用できる．

2.2 補完要素認定

補完要素の認定では，補完が必要な省略箇所に対する補完要素を認定する．単言語コーパスでは基本的にアナリストが各省略箇所に対する補完要素を人手で付与しなければならない．しかし，人手による付与作業においても，以下の点は考慮にいれる必要がある．

- 周辺に同じ構文意味構造の特徴（様相表現，用言の意味，接続表現等）を持つ省略箇所はグルーピングして，補完要素をタグ付けする際に一覧できるようにすべきである．
  周辺に同じ特徴を持つ省略箇所は同じタイプの補完要素を持つ傾向にある．これは，その特徴が補完要素を決定する手掛かりとなる要因となり得るからである．よって，アナリストは類似の省略箇所を一覧することで効率的に補完要素をタグ付けできる．
- 補完候補は，テキスト中や文章外の要素から容易に選択できるようにすべきである．
  省略箇所には3タイプの補完候補が考えられる：同じ文中の補完候補（文内），テキスト中の別の文中の補完候補（文間），テキスト中に明示されていない補完候補（文章外）．文内と文間の補完候補はテキスト中に明示されているため，これらの省略箇所に対する省略補完の条件としては，構文的位置や，文間関係（距離や文章構造中の意味的・相対的関係（例えば，補完候補が節のタイトル中にあり，省略がその節の本文中にある場合の関係）等）が有効である[7][8]．よって，テキスト中で同じ構文的位置や文間関係を周辺に持つ補完候補をグルーピングして一覧することで，省略箇所に対する補完要素を効率的かつ効果的にタグ付けできる．文章外では，補完候補が，書き手／話し手や読み手／聞き手のように限定された要素となる傾向にある[9]．よって，タグ付けの前にありえそうな補完候補を列挙し，省略箇所の対する補完候補を付与する際に列挙した候補から選択することで，文章外補完候補がより簡単に効率的に付与できる．

対訳コーパスでは，単言語文のみを活用した人手によるタグ付けに加えて，原言語中の省略箇所に対する対訳文中の翻訳表現も省略箇所の補完要素を認定するのに活用できる．例えば，日英対訳コーパスでは，主語や目的語は頻繁に省略されるが英語では必須要素となる場合が多い．よって，日本語文中の省略箇所と英訳文中のその訳語表現を対応付けすることで，省略箇所に対する補完要素が自動的に認定できる[13]．

2.3 ルール獲得不適文の排除[18]

コーパス中で以下のタイプの省略箇所および補完要素を含む文は，ルール獲得には適していない．

(a) 文の解析に失敗したもの

これは，対訳文対の文の解析において，解析処理系が誤った解析構造を生成してしまったために，その誤った解析構造情報を元に，規則を作成してしまう場合である．例えば，

(2) ぜったいにそうではない
"Certainly not."

では，日本語文には，「そうではない」の1カ所しか用言がないが，日本語解析の際に誤って，本来副詞である「ぜったい」も用言として認定してしまい，不適切な日本語構造が生成されてしまう場合がある．

(b) 意訳要約され忠実な対訳文対ではないもの

これは，文対としては，対訳関係にはあるが，要約されたり，意訳されたりして，文より細かな表現単位では対訳関係にある表現対の認定が困難である場合である．例えば，

(3) 国民は歓呼して彼を国王に迎えた
"The people acclaimed his king."

では，日本語文は2用言からなるが英語文では1用言を用いて意訳されている．これらの文対は対訳表現対の情報を利用してルール獲得する手法においては，獲得が困難な不適切な文対となる．

(a)に該当する不適文はタグ付けの際に'不適文'であるという情報を付与し，ルール作成の際に活用しない用の明示する必要がある．(b)に該当する省略に対しては，対訳コーパスであっても，補完要素を人手でタグ付けすることが望ましい．

2.4 ルール抽出

ルール抽出処理では，省略補完タグ付きコーパスから省略補完ルールを抽出する．ここでは，省略箇所と補完要素の周辺の構文意味構造の特徴を主な省略補完ルールの条件として用いる．

(a) 自動抽出

この処理では，省略補完ルールは省略補完タグ情報及び省略箇所と補完要素の周辺の特徴量を活用して，機械学習[14][15]または，統計処理[12]により自動的に抽出する．

(a) 人手抽出

この処理では，コーパス中の省略箇所をその構文的位置，タグ付けされた補完要素，省略箇所周辺の構文的意味的特徴によりグルーピングする．グルーピングした省略箇所に対して，同じ特徴下で同じ補完要素がどのくらいタグ付けされているかを検証することで，グルーピングされた省略箇所に対する補完ルールが抽出できる．

2.5 獲得ルールの適用と修正

2.4で獲得したルールは自然言語処理システムの省略補完処理中で，ルール獲得で活用したコーパスに対して適用する．省略箇所へのルールの適用結果を検討して，獲得ルールがこのコーパスにおいて適切であるか検討する．もし補完ルールに問題が発生したら，そのルール自体かそのルールの優先順位を修正する．

修正後のルールセットはその自然言語処理システムで同じコーパスに対して適用され，最適なルールセットが獲得できるまで同様の確認を繰り返す．

3. 省略補完ルール獲得環境の構成

2で述べた5種類のサブタスクを考慮にいれて，日本語単言語コーパス及び日英対訳コーパスから，自動及び人手で省略補完ルールを獲得するソフトウェア環境を実装した．図1は，この環境の構成図である．この環境では，まず対訳コーパス中の日英それぞれの文及び日本語単言語コーパス中の文を，日本語及び英語の解析系によってそれぞれ解析する．次に，コーパス中の各日本語文の省略箇所に対する補完要素を，対訳コーパスの日英解析結果をもとに自動認定する．しかし，単言語コーパスの場合には，省略箇所の構文的位置情報が付与された日本語解析結果のみが得られる．これら補完要素をタグ付けした／していない日本語解析結果は図のとおり「省略補完タグ付き日本語コーパス」として保存される．コーパス中の各省略箇所に対しては，必要に応じて，正確な補完要素を人手でタグ付けする．次に，タグ付けした情報に基づき，省略補完ルールを自動か人手で作成する．人手で作成する場合に

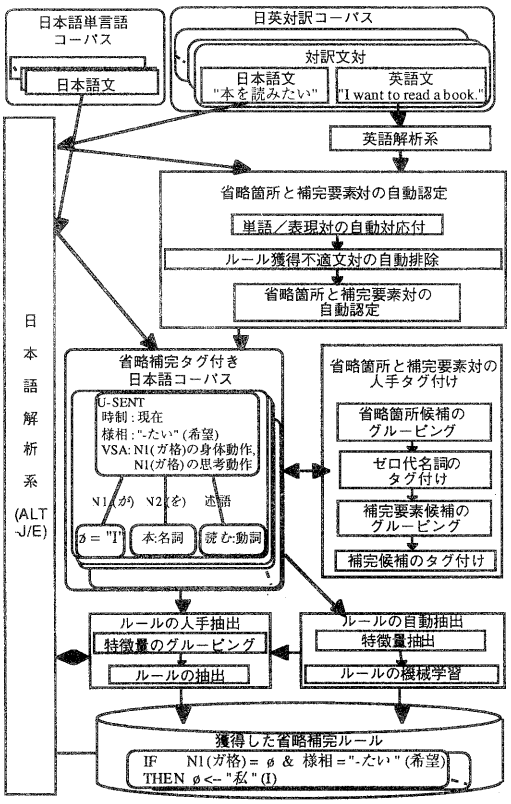

図1 省略補完ルール獲得環境の構成

は，正確なルールの獲得ができるが，作成コストが高くなる．それに対し，自動で作成する場合には，不正確なルールを獲得してしまう可能性がある．よって，信頼できるルールセットを獲得するためには，自動獲得ルールを再度人手でチェックし，必要に応じて修正することが望ましい．

次に，上記で獲得した省略補完ルールセットは日本語解析系で，ルール獲得対象となった同コーパス中の省略箇所に対して適用される．単言語及び対訳コーパス中の同じ文章をこの環境に再度入力し，ルールの獲得と修正を行う．この処理は，本環境により省略補完ルールが獲得できなくなるまでくりかえす．

本環境は，日英機械翻訳システム ALT-J/E[19]を活用して実装した．図1の環境を活用すれば，対訳文対から日本語省略箇所に対する英訳表現の情報も抽出できる．よって，この抽出結果は，日英機械翻訳システムにおける日本語省略箇所の翻訳ルールの抽出にも活用できる．なお，人手処理を効率化するために，本環境では，Netscape Navigator[1]や Internet Explorer[2]等の汎用的な WWW ブラウザのインタフェースを活用した．

[1] Netscape Communications 社の登録商標
[2] Microsoft 社の登録商標

個々の処理について次節以降で詳細に述べる．

3.1 日本語文と英語文の解析処理

コーパス中の日本語文と英語文は下記の手順により解析する．

3.1.1 日本語文解析

日本語文は，NTT が研究開発中の日英機械翻訳システム ALT-J/E[16][17]の形態素解析，構文意味解析処理系によって解析する．ALT-J/E では，日本語の構文意味構造は，直接翻訳する英文の構造と対応が付けられている．よって，日本語構造は，英語で訳出が必要となる省略箇所の位置情報と動詞による省略箇所への意味的制約の情報を含んでいる．日本語解析の結果，省略箇所に対して補完ルールが適用された場合，適用ルールの ID 番号が省略箇所にタグ付けされる．この情報は，既存のルールにより正確に省略補完できるか，どの省略個所が追加のルールを必要としているがを判断するのに活用する．例えば，日英対訳文対(4)の日本語文からは(5)のような日本語構文意味構造が生成される．

(4) (ϕ-が) 本を読みたい
"I want to read a book."

```
(5) u-sent-1
    時制 : 過去，完了相
    様相 : たい (希望)
    用言意味属性 : ガ格の身体動作，ガ格の思考動作
      ¦--- PRED : pred-1
      ¦              主動詞 : 読む (read)
      ¦--- CASE : case-1
      ¦       ¦    格関係 : 目的語
      ¦       ¦    助詞   : を
      ¦       ¦---- NP : np-1
      ¦               ¦-- N : 本
      ¦--- CASE : case-1
              ¦    格関係 : 主語
              ¦    助詞   : が
              ¦---- NP  : φ-1 (意味制約 : 人間)
```

3.1.2 英語文解析

英語文は，Brill の英語 tagger[20]及び英語構文解析ツール Link Grammar Parser[21]により解析する．この解析結果をもとに，ALT-J/E の内部英語構造に近い部分構文構造に変換する．例えば，日英対訳文(4)の英語文からは(6)のような英語部分構文構造が生成される．

```
(6) u-sent-1
      ¦--- PRED : pred-1
      ¦           "want" : verb, non-3rd, sing. present.
      ¦           "to"   : to
      ¦           "read" : verb, base from
      ¦--- CASE : case-1
      ¦       ¦    格関係 : subject
      ¦       ¦---- NP : np-1
      ¦               ¦-- N : "I" : personal pronoun
      ¦--- CASE : case-1
              ¦    格関係 : direct object
              ¦---- NP : np-2
                      ¦- "a"    : determiner
                      "book" : noun, singular or mass
```

3.2 省略箇所と補完要素の自動認定

日英対訳文対の解析結果から，日英の２種類の構造を比較して日英対訳単語／表現対を認定する．それから，2.3 で述べたルール獲得に適さない対訳文対を下記の条件に基づき自動的に排除する処理を行う[18].

- 日本語動詞が英語名詞と対訳関係にあると認定された日本語文の単文部分を除いて，日本語文の単文の数と英語文の単文の数が異なる場合．
- 日本語文の日英機械翻訳システムによる機械翻訳の結果，機械訳文中に翻訳できない日本語表現が残った場合．
- 英語文の構文解析系による解析の結果，解析できない部分が残った場合．

次に，日本語文中の省略箇所と英語文中のその英訳表現の対を人手作成した１０種類の規則に基づき自動認定する[12][13]．例えば，日英対訳文対(4)の日本語文中の主語（ガ格）の省略箇所からは，図１の省略補完タグ付き日本語コーパスの例のように，英語文中の主語の要素"I"が自動的に認定される．

3.3 省略箇所と補完要素の人手認定

日本語単言語コーパスでは，コーパス中の省略箇所向けの省略補完ルールを作成するアナリスト自身が，個々の省略箇所に対する補完要素を人手でタグ付けする必要がある．そこで効率的なタグ付けのために，本環境では日本語コーパス中の省略箇所に対する補完要素をタグ付けするツールを実現した．この処理では，日本語解析結果（3.1.1）を活用している．本処理の詳細を以下に示す．

3.3.1 省略箇所の人手認定

日英機械翻訳システムの日本語解析系では，明示的に英訳する必要のある日本語文中の省略箇所が，日本語構文意味構造中に明示的にタグ付けされる（例えば例(5)）．しかし，2.3 でも述べたとおり，その文の解析失敗が誤った省略箇所を生成してしまう可能性がある．よって，アナリストは解析結果中の省略箇所が本当の省略か否かをタグ付けする必要がある．効率的なタグ付けのために，本処理では，日本語文の解析結果は省略箇所の周辺の以下のような特徴に基づきグルーピングする．

(1) 省略箇所候補の構文的位置（ガ格，ヲ格等）
(2) 省略箇所候補周辺の構文意味構造（接続表現のタイプ，用言意味属性，省略箇所候補を含む単文の様相表現等）

省略箇所候補を構文的位置に基づきグルーピングした結果の本ツール画面を図２に示す．この図では，対象コーパス中で N1（ガ格）に省略箇所候補が最も多く

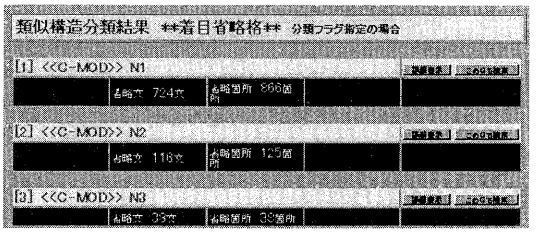

図2　省略箇所候補の構文位置によるグルーピング結果画面

出現（724 文中に 866 箇所）し，次いで N2（ヲ格）に多く出現（116 文中に 125 箇所）したことを示している．

3.3.2　補完要素のタグ付け

日本語文の省略箇所の認定後，個々の省略箇所に対する補完要素をタグ付けする．省略箇所の人手認定と同様に，省略箇所はその周辺の特徴量によってグルーピングする．次に個々の省略箇所の補完要素を効率的にタグ付けするために，文内及び文間の補完要素候補をその周辺の以下の特徴量によってグルーピングする．

(1) 補完要素候補の構文的位置

(2) 補完要素候補周辺の構文意味構造

(3) 省略箇所の単文と補完要素候補の単文との構文的な関係（両者の単文が直接的に接続表現で繋がっている等）（文内のみ）

(4) 省略箇所の文と補完候補の文の文章構造的（又は距離的）な関係（省略箇所の文が補完候補の文の次にある等）（文間のみ）

(3)の特徴量に関しては，まず，省略箇所を含む単文の構文構造を分類し，その後，同じタイプの構文構造を持つ文を分類する．(4)の特徴量に関しては，まず，対象コーパスでよく現れる，省略箇所と補完要素間の位置的関係を事前に調査する．例えば，新聞記事では，記事の第 1 文中にそれ以降の文の省略箇所の補完要素が多く存在することが知られている[7]．このように，頻出する補完要素と省略箇所の位置的関係をルール作成分野向きにリストアップしておき，その位置的関係の文から補完要素を選択することで，文間でも効率的なタグ付けが可能となる．文章外の補完要素をタグ付けする際には，“I / we”や“you”, “it”のように頻出する文章外補完要素候補を事前にリストアップし，そこから正しい補完要素を選択する．省略箇所に対する補完要素がタグ付けされたら，それ以外の補完要素候補は“誤った補完要素”としてグルーピング結果の画面に表示する．

同じ特徴を持つ補完要素候補をグルーピングすることにより，アナリストは同じタイプの文脈中の省略現象を同時に参照できるので効率的なタグ付けが可能となる．さらに，高頻度文脈中の省略から低頻度文脈中の省略に順にタグ付けすることにより，アナリストは早い段階から効率的なタグ付けが可能となる．

3.4　省略補完ルールの自動抽出

省略補完ルールの自動抽出では，まず，省略箇所と補完要素の周辺の構文的・意味的特徴量を省略補完タグ付けされた日本語文解析構造から抽出する．以下に示すような，その有効性が実証されている特徴量[7][8][9]を自動抽出する省略補完ルールの条件として活用する．

(1) 用言意味属性[5]

(2) 様相表現のタイプ

(3) 省略箇所の単文と補完要素の単文を繋ぐ接続表現のタイプ

(4) 省略箇所の単文と補完要素の単文の構文的位置関係（文内のみ）

(5) 省略箇所の文と補完要素の文の文章構造的関係（文間のみ）

省略補完ルールは決定木学習プログラム C5.0[22]により自動的に抽出する．その後，抽出ルールは ALT-J/E の省略補完ルールの書式に変換される．

3.5　省略補完ルールの人手抽出

人手作業を介して信頼性の高い省略補完ルールを抽出するために，省略補完タグ付き日本語文解析結果から人手でルールを作成するツールを本環境中に実装した．3.3 で述べた方法と同様に，省略箇所と補完要素の周辺の 3.4 で示した 5 種類の特徴量に基づきグルーピングを行い，個々のグループの頻度でソートしてその結果を画面表示する．このようなグルーピングにより，ルール抽出の最初の段階から，カバレージの広いルールを効率的に抽出することができる．また，同じタイプの文脈中の省略箇所を考慮にいれたルール抽出も可能となる．さらに，個々のルールの適用省略箇所数とその中で正しく補完要素が認定できた省略箇所数の比を検証することにより，ルール抽出の段階でも個々の抽出ルールの信頼度が推定できる．抽出ルールは ALT-J/E の省略補完ルールに追加する前に，ルール間の包含関係を検証し，全ルールに対する獲得ルールの適切な優先順位を設定する．

4．評価

本稿で提案した環境の構成要素のうち，省略補完ルールの自動獲得に関しては，その認定精度が[12][13]において既に報告されている．この評価結果によれば，日本語省略箇所とその英訳表現の対の自動認定精度は，

クローズドテストで 98.4%，オープンテストでも 94% と高い値が達成できた．また，自動獲得ルールの省略補完性能の評価によると，文章外の省略補完ルールについては，クローズドテストで 99.0%，オープンテストで 85%の精度が得られたことが報告されている．よって，ここでは，日本語単言語コーパスから提案した環境を用いて人手でルールを獲得する処理の評価を行う．3 で述べたとおり，提案手法の人手ルール獲得での効率性は，グルーピング機能に強く依存する．よって，本評価では，グルーピング機能の有無にともなうルール獲得の効率の変化を検証した．

4.1 人手ルール獲得の評価方法

提案環境を活用した人手によるルール獲得の有効性と効率性を日英機械翻訳用評価例文中の日本語文(3719 文)[23]を用いて評価した．ALT-J/E の省略補完処理を熟知したアナリストが 3 で述べたシステム（SUN Spark Enterprise 3000 中にインストール）を活用して省略補完ルールを下記の 2 種類の手順で抽出した．

(a) グルーピング機能を活用したルール獲得

グルーピング，タグ付け及びルール抽出は以下の手順で行った．

[Step-1] 省略箇所候補をその構文的位置に基づきグルーピングする．ここでは，最も頻度の高い N1（ガ格)の省略箇所を選択する：724 文中の 866 箇所（図 2）．

[Step-2] 選択した省略箇所候補は再度その構文構造に基づきグルーピングする．ここで最も頻度の高い構文構造である，省略箇所候補の単文が接続表現を介して直接別の単文と繋がっている構造を選択する：315 箇所．

[Step-3] 本当の省略箇所であるか否かを人手判定しタグ付けする：315 箇所中 285 箇所．

[Step-4] 選択した省略箇所に対して接続表現のタイプによってグルーピングした後，個々の省略箇所に対する補完要素をタグ付けする．

[Step-5] 285 箇所の省略のタグ付け結果をもとに人手により最初の 5 ルールを獲得し，5 ルールの獲得に要した時間を計測する．

(b) グルーピング機能を活用しないルール獲得

解析結果で省略箇所が含まれる文に対して，グルーピング機能を活用せずに，1 文 1 文人手分析して省略補完ルールを獲得する．この作業は，(a)においてグルーピング機能を活用して 5 ルールを獲得するまでに要した時間だけ続ける．

2 種類のルール獲得結果は，(1) 省略箇所に対する補完要素のタグ付けが効率的に早く行えたか，(2) 獲得されたルールによって何箇所の省略が正しく補完できたか，の 2 種類の観点から検討し評価した．

4.2 評価結果

本評価実験の結果を表 1 に示す．この表が示すとおり，省略箇所と補完要素のタグ付けは，テスト(a)の方がテスト(b)より効率的に行うことが出来た（グルーピング機能を活用した場合（テスト(a)）は，省略箇所タグ付け 1.1 分/箇所，補完要素タグ付け 1.7 分/箇所に対し，グルーピング機能を活用しなかった場合（テスト(b)）は，省略箇所タグ付け 2.5 分/箇所，補完要素タグ付け 2.0 分/箇所）．さらに，ルールの抽出時間と獲得ルールの適用時間に関しても，テスト(a)の方がテスト(b)より早く行えることが分かった（テスト(a)ではルール抽出 1.1 分/箇所，ルール適用 2.2 分/箇所に対し，テスト(b)ではルール抽出 6.0 分/箇所，ルール適用 10.0 分/箇所）．これら結果から，タグ付け情報が付与されたグルーピング結果の情報を活用した場合は，グループ内の省略箇所に基づき獲得したルールの認定精度をルール適用前に予測できるので，カバレージの広いルールを効率的に獲得するのに有益であることが分かった．

獲得ルールに補完精度に関しては，テスト(b)のルールの方がテスト(a)よりも正確に補完することができた（テスト(a)が 93%に対しテスト(b)が 100%）．このテスト(a)のルールにより補完に成功しなかった 5 箇所は，アナリストはルールを抽出する時点で，グルーピング後の正誤情報から問題が発生する箇所であることが認識できていた．よって，グルーピング機能を活用した場合には，それらの問題省略箇所に対しても，グルーピング結果を参照してルールの条件をより詳細化することで正確なルールを容易に獲得することかできることがわかった．

表 1 ルールの人手獲得の必要時間と精度
（1 省略箇所当りの作業時間[分]を括弧内に示す）

グルーピング機能		使用 (a)	未使用 (b)
獲得ルール数		5	51
作業時間[分]	省略箇所タグ付け	332 (1.1)	128 (2.5)
	補完要素タグ付け	482 (1.7)	102 (2.0)
	ルール抽出	77 (1.1)	306 (6.0)
	ルール適用	128 (2.0)	510 (10.0)
	計	1049 (6.0)	1046 (20.5)
ルール適用省略箇所数		72	51
補完に成功した省略箇所数		67 (93%)	51 (100%)

5．まとめ

本稿では，単言語コーパス及び対訳コーパスをから効率的に省略補完ルールを獲得する実用的なソフトウェア環境を提案した．本環境を活用した省略補完ルールの作成評価により，提案手法のグルーピング機能を有効に活用することで，単言語コーパスからの人手作業による場合でも，省略補完ルールを効率的に獲得できることがわかった．今後は，提案手法を単言語・対訳の両コーパスでの有効性を厳密評価をすると共に，自動処理と人手処理の最適な組合せ手法についてもさらに検討を行っていきたい．

参考文献

[1] Kameyama,M.: A Property-sharing Constraint in Centering, Proc. of ACL (1986).

[2] Walker, M. et al.: Centering in Japanese Discourse, Proc. of COLING'90 (1990).

[3] Yoshimoto, K. : Identifying Zero Pronouns in Japanese Dialogue, Proc. of COLING'88 (1988).

[4] 堂坂：語用論的条件の解釈に基づく日本語ゼロ代名詞の指示対象同定，情報処理学会論文誌，Vol.35 No.5 (1994).

[5] 中岩,池原：日英の構文的対応関係に着目した日本語用言意味属性の分類，情報処理学会論文誌，Vol.38 No.2 (1997).

[6] 池原ら編: 日本語語彙大系，岩波書店 (1997).

[7] 中岩,池原：日英翻訳システムにおける用言意味属性を用いたゼロ代名詞照応解析，情報処理学会論文誌,Vol.34 No.8 (1993).

[8] 中岩,池原：語用論的意味論的制約を用いた日本語ゼロ代名詞の文内照応解析，自然言語処理，Vol. 3 No.4 (1996).

[9] 中岩，白井，池原：日英機械翻訳における語用論的・意味論的制約を用いたゼロ代名詞の文章外照応解析，情報処理学会論文誌，Vol.38, No.11 (1997).

[10] Nasukawa, T.: Full-text processing: improving a practical NLP system based on surface information within the context, Proc. of COLING-96 (1996).

[11] 村田，長尾: 用例や表層表現を用いた日本語文章中の指示詞・代名詞・ゼロ代名詞の指示対象の推定，自然言語処理，Vol.4, No.1 (1997).

[12] Nakaiwa, H.: Automatic Extraction of Rules for Anaphora Resolution of Japanese Zero Pronouns from Aligned Sentence Paris, Proc. of ACL/EACL-97 workshop on Operational Factors in Practical, Robust, Anaphora Resolution for Unrestricted Texts (1997).

[13] 中岩：日英対訳コーパス中のゼロ代名詞とその指示対象の自動認定，情報処理学会研究報告，NL-123-5 (1998).

[14] Aone, C. and Scott W. Bennett : Automated Acquisition of Anaphora Resolution Strategies, Working Notes of AAAI Spring Symposium Series, Empirical Methods in Discourse Interpretation and Generation (1995).

[15] Yamamoto, K. and E. Sumita : Feasibility Study for Ellipsis Resolution in Dialogues by Machine-Learning Technique, Proc. of COLING-ACL-98 (1998).

[16] Hasida, K. : Global Document Annotation (GDA), http://www.etl.go.jp/etl/nl/GDA/ (2000).

[17] Aone, C. and Scott W. Bennett : Discourse Tagging Tool and Discourse-Tagged Multilingual Corpora, Proc. of the Intarnational Workshop on Sharable Natural Language Resources (1994).

[18] 中岩：対訳コーパス中の規則獲得不適文対の自動認定,情報処理学会第 57 回全国大会,5R-8 (1998).

[19] 八巻他：特集論文 日英機械翻訳技術，NTT R&D Vol. 46 No. 12 (1997).

[20] Brill, E. : A simple rule-based part of speech tagger, Proc. of ANLP'92 (1992).

[21] Sleator , D. and D. Temperley : Parsing English with a Link Grammar, CMU Computer Science Technical Report, CMU-CS-91-196 (1991).

[22] Quinlan, J. R. : http://www.rulequest.com/ (1998).

[23] 池原,白井,小倉 :言語表現体系の違いに着目した日英機械翻訳機能試験項目の構成，人工知能学会誌論文，Vol.9, No.4 (1994).