認識構造を抽出する英語文パーザ

加藤 尚之　　宮崎 正弘

新潟大学大学院自然科学研究科

1 はじめに

計算機による自然言語処理において、従来の言語の形式のみの把握だけでは十分とはいえない。より高度な自然言語の意味理解の実現には、自然言語の本質についての考察が必要である。そのため、人間の言語活動である「対象⇔認識⇔表現」という過程的構造に着目した時枝誠記による言語過程説、その言語過程説を発展的に継承した三浦つとむの考え [1] に基づいて、言語の本質について検討し、計算機上での処理の方法についても検討した。また、計算機上で英語文の本質的な意味理解を行うために、三浦つとむの説を英語に適用した宮下眞二の英語文法の考え方 [2] を取り入れ、言語の統語構造に相当する表現構造と認識された対象世界を把握する認識構造を抽出する英語文パーザを試作した。既に一般化ＬＲ法を Prolog 上に実現したＳＧＬＲパーザ [3] を拡張した拡張型ＳＧＬＲパーザＳＧＬＲ-plus[4] 上に英語文パーザ [5] が試作されている。このパーザを発展的に拡張し、中学レベルの英語文を網羅的に解析可能とした。

2 言語過程説

時枝誠記は、人間の言語活動である「対象⇔認識⇔表現」に着目し言語過程説を唱えた。その言語過程説を発展的に継承した三浦つとむ [1] は、言語について次のように述べている。

人間の言語活動には言語生成と言語理解の２つがある。まず言語生成の場合、表現しようとする対象の客観的な姿を概念として表す。かつ、その対象に対し主観的で様々な判断を重ね合わせる。これにより人間の頭の中に認識構造を形成する。そしてその認識構造から発話、記述するなりして表現しようとするが、そこには必ず社会的な約束である言語規範（文法、辞書、…）で拘束される。このようにして人間は言語を表現する。また、逆にある言語表現を言語規範に基づいて聞いたり読んだりして認識構造を頭の中に形成する。そしてその認識構造から主体の持つ世界知識と重ね合わせ推論することによって言語表現は理解される。図１はこの様子を簡単に表したものである。

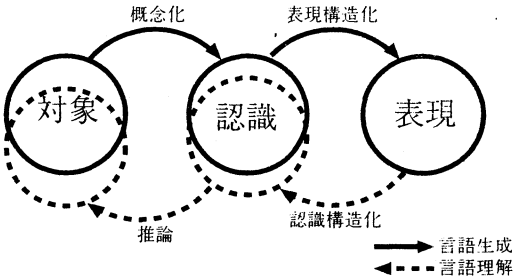

図 1: 言語の過程的構造

3 認識構造構築の過程

3.1 客体認識と主体判断

人間の言語表現は文字や音声によって表される。それらは文として存在し、表現と呼ばれる。それが各言語において文法上どのような構造かを考えたものが表現構造である。その構造は「対象を客観的に捉えたものを概念として表す客体的表現」と「それに対する人間の主観的な判断、感情、客体認識の中の関係等を表す主体的表現」に分かれる。この分割作業は人間が言語規範によって聞いたり読んだりする過程で行われ、それを本能的に噛み砕いて客体認識と主体判断に変換する。そして認識構造を形成する。その様子を図２に示す。

A Experimental English Parser for Extracting Speaker's Recognition Structure
Naoyuki Kato , Masahiro Miyazaki
Niigata University

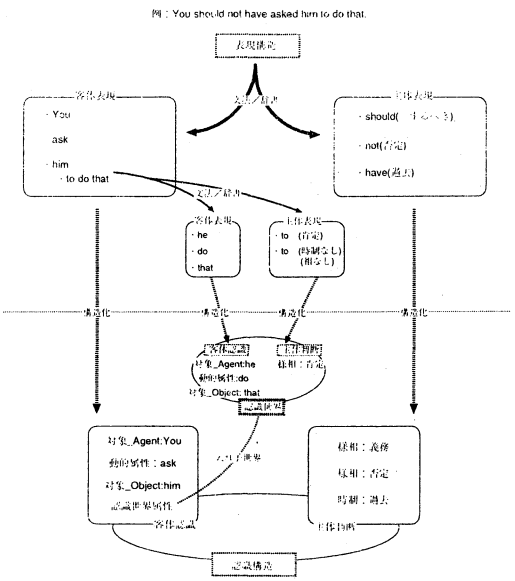

図 2: 表現構造から認識構造へのながれ

4 認識構造

【客体認識】…対象を客観的にとらえたものを概念として表したもの
【主体判断】…客体認識に対する人間の主観的な判断、感情を表したもの
認識構造はこれら 2 つを重ね合わせて形成される。認識構造の構成要素を以下に示す。

4.1 客体認識を構成する要素

・動的属性
　実体の動作を表す
・静的属性
　対象の状態を表す
・対象__Agent
　話者が注目している、また動作を行う対象の実体
・対象__Object
　動的属性の効果が及ぼす目標
・推移__Agent
　動的属性の効果が及ぼした後の対象の姿
・受益者
　動的属性の効果が及ぼした後の対象__Object の移動先
・～格
　空間、時間、道具、方向…などの情報（in 格、at 格など；前置詞句，ϕ 格；副詞）
・従属世界属性
　関係代名詞節／関係副詞節／従属接続節などの埋め込み文にあたる入れ子で存在する世界

4.2 主体判断を構成する要素

・判断
　客体認識に対する表現主体の直接的な意思判断（肯定／否定）。
・時制
　表現主体と対象との時間的関係の認識を直接的に表現した主体判断。
・態
　主語が目的語に働きかけているか、逆に働きかけているかといった実体間における
　動的属性の波及の仕方を表した客体表現。
・様相
　客体認識に対する表現主体の直接的な判断（疑問／命令など）。
　表現主体の判断を直接的に表現した主体表現（推量／意思／宣言／義務／許可／丁寧など）
・相
　実体のもつ動的属性が完了しているか、未完了か、進行中かなどを表す客体表現。
・実体把握
　客体表現部分の構成要素が表す主体表現であり実体の捉え方を表す（総称、特定、不特定など）

5 英語文パーザの作成

以上の検討に基づき、計算機による言語理解に向けた英語文の表現構造を抽出する英語文パーザのプロトタイプを構築する。

5.1 プロトタイプの構成

パーザ本体としては SGLR（A Seaquential Generalized LR Parser; 逐次型一般化 LR）パーザ [4] を用いる。この SGLR パーザは、横型探索をベースとした富田法 [3] を Prolog 上に記述したもので、先読み情報を利用することにより無駄のない高速な統語解析を実現する。また今回、関係代名詞節における痕跡の追跡、正規表現の記述、グラフ構造化などが可能となった SGLR の拡張版である SGLR-plus を使用した。

パーザで用いる文法は、補強 CFG（Context-Free Grammer；文脈自由文法）を Prolog 上で実現するために DCG（Definite Clause Grammer; 確定節文法）形式で記述する。この DCG 形式で記述された文法は最終的にトランスレータによって Prolog プログラムに変換される。Prolog プロ

グラムに変換して実行することにより、Prologの基本計算機構をそのまま利用した自然言語処理を行なうことができる。DCGの文法形式は次のような特徴がある。

- 非終端記号を述語とみなし、それに任意この引数を持たせることができる。
- 記号"−−>"の右側の任意の場所にDCGの補強項と呼ばれる中括弧'{}'で括られた任意のゴール文を挿入することができる。
- 記号"−−>"の右には、終端記号と非終端記号とが混在した任意の長さの記号列の記述ができる。つまり、CFGの一般形が記述できる。
- 補強項の中に、非終端記号に対応する述語の中で使われている引数を使うことができるので、統語的な制約が容易に記述できる。またそれにより、文法規則の精度をあげ統語解析速度をあげることも可能となってくる。

5.2 辞書

EDR英語辞書を辞書引きプログラムに組み込んだことにより、英和辞典に通常のっている単語を含む文の解析が可能となっている。全ての情報を、常に引数として渡し続けるわけにはいかないので、語形変化情報／文法情報／文型情報に分割し、別々のファイルにDCG形式で書き出し、必要時のみ補強項から呼び出すという手法をとり、解析必要最低限の情報に関してはそのままパーザに渡すことにする。また今回はEDR英語辞書に収録されていない未知語に関して、検索件数が0件の場合は名詞の情報を与えるプログラムを、辞書引きプログラムに組み込んだ。これにより、未知語を含んだ英語文でも解析可能となり、解析可能文の幅が広がった。その辞書引きの様子を図3に示す。

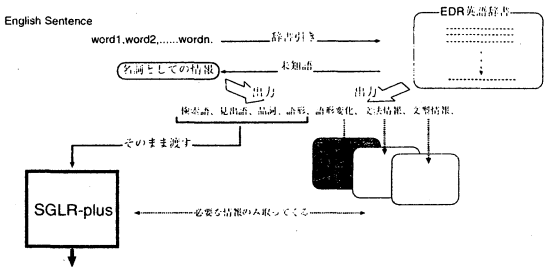

図 3: 辞書引きの流れ

6 定量評価

6.1 評価方法

「日英機械翻訳機能試験文集」(全6200文)[6]における500文を用いて評価実験を行った。基本的には認識構造が正しく出力されるかどうかを判断し、その結果を表1に記す。また、機能試験文の特性を表2に示す。なお、1つの文に対して出た結果に多義が含まれていても、正解の木が出力されていれば良いものとする。

表 1: 正解率

評価件数	解析可能件数	正解率(%)
500	450	90.0

表 2: 機能試験文の特性

総単語数	最大単語長	最小単語長	平均単語長
3331	19	2	6.662

6.2 多義曖昧性

1つの文を解析するにあたって、解析木が一意に決まらない場合が多くある。それが構造的な多義である。多義には様々な原因があり、一番の理由として挙げることができるのは単語の多品詞性の問題がある。これに伴い、文法の曖昧性が生じたりする。多義の数と解析件数の関係を表3に示す。

表 3: 多義数内訳

最大多義数	最小多義数	平均多義数
146	1	5.89

また、縦軸を多義数、横軸を単語長とした単語長と多義数の関係を図4に、多義数と累積割合を図5に示す。図6は解析機構出力する実際の解析例である。

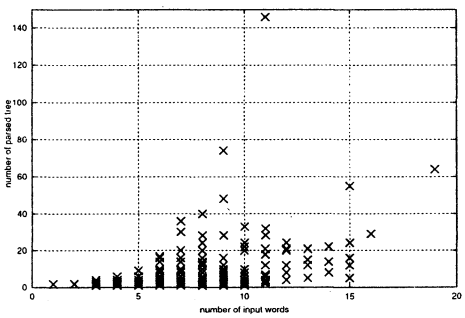

図 4: 単語長と多義数の関係

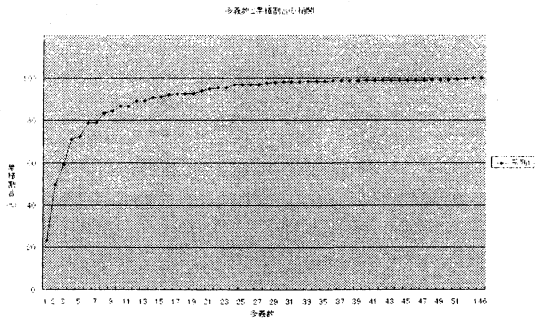

図 5: 単語長と多義数の関係

```
認識世界
 |- 客体認識
 |    |- 動的属性
 |    |    |- 概念 -- < get >
 |    |    |- 属性 -- < up >
 |    |- 対象_Agent
 |    |    |- 把握 -- < 特定 >
 |    |    |- 概念 -- < boy >
 |    |    |- 認識世界属性
 |    |    |    |- 客体認識
 |    |    |    |    |- 動的属性
 |    |    |    |    |    |- 概念 -- <meet>
 |    |    |    |    |- 対象_Agent
 |    |    |    |    |    |- 概念 -- <she>
 |    |    |    |    |    |- 深層格 --<動作主>
 |    |    |    |    |- 対象_Object
 |    |    |    |    |    |-把握--< 特定 >
 |    |    |    |    |    |-概念 -- < boy >
 |    |    |    |    |    |-深層格--< 対象 >
 |    |    |    |- 主体判断
 |    |    |    |    |- 判断 -- < 肯定 >
 |    |    |    |    |- 時制 -- < 過去 >
 |    |    |    |    |- 態 -- < 能動 >
 |    |    |    |    |- 様相 -- < φ >
 |    |    |    |    |- 相 -- < φ >
 |    |    |-深層格 --<動作主>
 |    |- for__格
 |    |    |- 把握 -- < 特定 >
 |    |    |- 概念 -- < park >
 |    |    |- 深層格 -- < 原因 >
 |    |- φ格
 |    |    |- 概念 -- < early >
 |    |    |- 深層格 -- < 空間 >
 |- 主体判断
      |- 判断 -- < 否定 >
      |- 時制 -- < 現在 >
      |- 態 -- < 能動 >
      |- 様相 -- < can >
      |- 相 -- < φ >
```

図 6 ;「the boy that she met cannot get up early for his sickness.」の認識構造

7 おわりに

本稿では、三浦の言語モデルに基づき、言語表現の持つ本質的な意味理解の実現を目指し、英語文の統語構造を介して、どのように認識構造を構築していくかを検討し、認識構造を抽出する英語文パーザを試作した。現段階では単文、複文、重文、前置詞句や副詞句を含んだ文、句動詞などが解析できる。また、評価実験を行い、解析機構の精度と多義による曖昧性を検証した。今後さらに、慣用表現や解析木の多義絞り込みの問題を検討する必要がある。また、本パーザを認識構造を介して多言語間での機械翻訳に有用なものへとしていく予定である。

参考文献

[1] 三浦つとむ：日本語とはどういう言語か, 講談社学術文庫 (1976)

[2] 宮下眞二：英語とはどういう言語か、季節社 (1985)

[3] 沼崎、田中：SGLR：逐次型一般化 LR パーザの Prolog による実現、情報処理学会論文誌、Vol.32,No.3,pp.396-403(1991)

[4] 五百川、宮崎：痕跡処理のための逐次型一般化 LR パーザ SGLR の拡張、言語処理学会第 4 回年次大会発表論文、pp.314-317(1998)

[5] 高草木伸高：ＳＧＬＲ-plus による話者の対象認識を抽出する英語文パーザの試作、情報処理学会第５８回全国大会講演論文集（２）pp.71-72(1999)

[6] 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、vol.9,No.4,pp.569-579(1994)