LAX 形態素辞書の記述形式と辞書記述の改良

白石智子　久保幸弘†
(財) 日本情報処理開発協会　†(財) 新世代コンピュータ技術開発機構

1　はじめに

ICOTでは、自然言語処理システムの構築に必要な文解析・文生成などのモジュールの開発環境をLTB(Language Tool Box)としてまとめている。形態素意味解析システムLAX[1]はこのLTBのツールの一つで、通常の日本語の文章から、文節の並びとその意味構造を解析するプログラムの開発環境である。

LAXの形態素意味解析プログラムでは、LAX辞書に書かれた接続規則に従って文の形態素解析を行い、さらに意味構成規則を用いて文節内意味の構成を行う。従来のLAX辞書の接続情報の記述で、若干の問題点はあるが満足のゆく結果を得ている[2]。現在、引き続き文節意味構成の記述実験を進めているが、その結果従来の接続情報の、特に活用に関する記述に改良すべき点が発見された。その改良を行った所、従来の問題点が解消されると共に全体の記述量が減少し、意味構成規則を記述する上で見通しの良い辞書の体系になったので報告する。

2　記述形式の改良

LAX辞書の文法は森岡の形態素論[3]に依拠している。記述の単位は形態素であり語基、接辞、助辞の3つに分類されている。記述の形式は各形態素に対して自分自身の識別子である形態素識別子と、接続しうる形態素の識別子の集合である接続識別子リストと、語の意味を構成するための意味構成規則が記述される[4]。従来

```
<< 旧記述形式 >>

begin(カテゴリ名).
  表層1  :: 形態素識別子
         && 接続識別子リスト
         ## 意味構成規則
  表層2  :: 形態素識別子
         && 接続識別子リスト
         ## 意味構成規則
        :
end(カテゴリ名).
```

```
<< 新記述形式 >>

cat(カテゴリ名)
    ::   形態素識別子
    &&   接続識別子リスト
    ##   意味構成規則
    begin(カテゴリ名).
    表層1, 表層2, … 表層N.
end(カテゴリ名).
```

図 1: 形態素の記述形式

のLAX辞書では図 1の旧記述形式のように、1語1語の形態素ごとの記述を行っていた。この方法は辞書開発の初期の接続素性の洗い出しには役立ったが、エントリの増加に伴い辞書記述量が増大すると、辞書全体が統一的に記述できなくなるという問題があった。そこで、同一の意味構成規則を持ち、かつ同一の接続識別子リストを持つ形態素の集まりを一つのカテゴリとみなし、カテゴリ毎に接続規則および意味規則を記述するようにシンタクスの改良を行った(図 1の新記述形式を参照)。その結果、辞書記述量が減少し、また形態素を追加する際に接続規則、意味規則の記述をする必要がなくなり、大規模辞書の開発が容易となった。

3　辞書記述の改良

従来のLAX辞書において活用語は、強変化活用助辞をはじめとする14種の活用語の型に分かれ、更に12種の活用形に分類されていた(図 2を参照)。この分類に

図 2: 従来の活用語体系

よる活用表をウ系強変化のカ / タ / ラ / ワ行についての活用を例にあげてみる(表 1を参照)。このように分類された形態素を分析すると、活用形ごとに非常に似通った体系であることがわかる。また、同一の活用形で同一の表層であるにもかかわらず、活用語の型が異なるために別々のカテゴリに登録されている語も幾つかある。このことから従来の分類法(まず活用語の型に分け、次に活用形ごとに分ける方法)は、見直す必要がある。活用語尾は、活用語の何の型から派生したかに関係なく、つま

表 1: ウ系強変化活用形

	カ活用	タ活用	ラ活用	ワ活用
現在形	く	つ	る	う
完了形	いた	った	った	った
現在推量形	こう	とう	ろう	おう
完了推量形	いたろう	ったろう	ったろう	ったろう
否定推量形	くまい	つまい	るまい	うまい
命令形	け	て	れ	え
現在条件形	けば	てば	れば	えば
完了条件形	いたら	ったら	ったら	ったら
並列形	いたり	ったり	ったり	ったり
現在中立形	き	ち	り	い
完了中立形	いて	って	って	って

Improvement of LAX Morphological Dictionary
Tomoko SHIRAISHI　Yukihiro KUBO †
JIPDEC　† ICOT

LAX形態素辞書の記述形式と辞書記述の改良

白石智子　久保幸弘†
(財) 日本情報処理開発協会　†(財) 新世代コンピュータ技術開発機構

1　はじめに

ICOTでは、自然言語処理システムの構築に必要な文解析・文生成などのモジュールの開発環境をLTB(Language Tool Box)としてまとめている。形態素意味解析システムLAX[1]はこのLTBのツールの一つで、通常の日本語の文章から、文節の並びとその意味構造を解析するプログラムの開発環境である。

LAXの形態素意味解析プログラムでは、LAX辞書に書かれた接続規則に従って文の形態素解析を行い、さらに意味構成規則を用いて文節内意味の構成を行う。従来のLAX辞書の接続情報の記述で、若干の問題点はあるが満足のゆく結果を得ている[2]。現在、引き続き文節意味構成の記述実験を進めているが、その結果従来の接続情報の、特に活用に関する記述に改良すべき点が発見された。その改良を行った所、従来の問題点が解消されると共に全体の記述量が減少し、意味構成規則を記述する上で見通しの良い辞書の体系になったので報告する。

2　記述形式の改良

LAX辞書の文法は森岡の形態素論[3]に依拠している。記述の単位は形態素であり語基、接辞、助辞の3つに分類されている。記述の形式は各形態素に対して自分自身の識別子である形態素識別子と、接続しうる形態素の識別子の集合である接続識別子リストと、語の意味を構成するための意味構成規則が記述される[4]。従来

<< 旧記述形式 >>

```
begin(カテゴリ名).
 表層1  :: 形態素識別子
        && 接続識別子リスト
        $$ 意味構成規則
 表層2  :: 形態素識別子
        && 接続識別子リスト
        $$ 意味構成規則
        :
end(カテゴリ名).
```

<< 新記述形式 >>

```
cat(カテゴリ名)
  ::   形態素識別子
  &&   接続識別子リスト
  $$   意味構成規則
  begin(カテゴリ名).
  表層1, 表層2, … 表層n.
end(カテゴリ名).
```

図 1: 形態素の記述形式

のLAX辞書では図 1の旧記述形式のように、1語1語の形態素ごとの記述を行っていた。この方法は辞書開発の初期の接続素性の洗い出しには役立ったが、エントリの増加に伴い辞書記述量が増大すると、辞書全体が統一的に記述できなくなるという問題があった。そこで、同一の意味構成規則を持ち、かつ同一の接続識別子リストを持つ形態素の集まりを一つのカテゴリとみなし、カテゴリ毎に接続規則および意味規則を記述するようにシンタクスの改良を行った(図 1の新記述形式を参照)。その結果、辞書記述量が減少し、また形態素を追加する際に接続規則、意味規則の記述をする必要がなくなり、大規模辞書の開発が容易となった。

3　辞書記述の改良

従来のLAX辞書において活用語は、強変化活用助辞をはじめとする14種の活用語の型に分かれ、更に12種の活用形に分類されていた(図 2を参照)。この分類に

図 2: 従来の活用語体系

よる活用表をウ系強変化のカ / タ / ラ / ワ行についての活用を例にあげてみる(表 1を参照)。このように分類された形態素を分析すると、活用形ごとに非常に似通った体系であることがわかる。また、同一の活用形で同一の表層であるにもかかわらず、活用語の型が異なるために別々のカテゴリに登録されている語も幾つかある。このことから従来の分類法(まず活用語の型に分け、次に活用形ごとに分ける方法)は、見直す必要がある。活用語尾は、活用語の何の型から派生したかに関係なく、つま

表 1: ウ系強変化活用形

	カ活用	タ活用	ラ活用	ワ活用
現在形	く	つ	る	う
完了形	いた	った	った	った
現在推量形	こう	とう	ろう	おう
完了推量形	いたろう	ったろう	ったろう	ったろう
否定推量形	くまい	つまい	るまい	うまい
命令形	け	て	れ	え
現在条件形	けば	てば	れば	えば
完了条件形	いたら	ったら	ったら	ったら
非列形	いたり	ったり	ったり	ったり
現在中立形	き	ち	り	い
完了中立形	いて	って	って	って

Improvement of LAX Morphological Dictionary
Tomoko SHIRAISHI　Yukihiro KUBO †
JIPDEC　† ICOT