ＬＴＢ－形態素解析システムＬＡＸの開発環境

4C-4

久保 幸弘　＊妙泉 正隆　佐野 洋　赤坂 宏二　杉村 領一

（財）新世代コンピュータ技術開発機構　＊（財）日本情報処理開発協会

１．はじめに

本稿では，汎用日本語処理系ＬＴＢにおける形態素解析システムＬＡＸの開発環境について述べる．ＬＡＸでは正規言語と等価な形態素文法[1]を形態素辞書の形で記述し，それをトランスレートして形態素解析プログラムを得る[2]．本システムでは逐次型ＬＡＸアルゴリズムをサポートし，この形態素辞書の開発と文法記述の効率的なデバッグを行うためのツールとして，それぞれ文法辞書エディタとインスペクタを備えている．これらのツールを用いて開発された形態素解析プログラムはＬＴＢの一環としてＳＡＸの前処理部的な位置付けとなるが，単独でも使用可能なため，日本語インタフェースの入力部としての応用も容易である．

２．形態素解析プログラム

ＬＡＸで開発される形態素解析プログラムは図１の様に，変換された形態素辞書と解析エンジンからなる．形態素辞書は複数のクラス構成となっている．解析は二つのフェーズからなる．先ず，文章が入力されると解析エンジンは形態素辞書を用いて入力文の形態素解析を行い，文節候補を求める．このとき解析は決定的に行い，解析結果のオルタナティブを縮退されたデータ構造として一度に出力する．次に，この結果と各形態素の持つ構文情報を用いて，各文節候補の構文情報を決定する．これがＬＡＸにおける形態素解析結果と呼ばれるものであり，構文解析システムＳＡＸ[3,4]の入力データとなる．

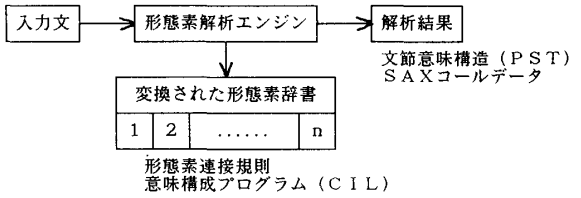

図１　形態素解析プログラムの構成

３．形態素辞書記述形式

図２に形態素辞書記述形式を示す．大きく分けて連接規則の記述部と意味情報の記述部とに分けられる．後方連接素性には複数の状態遷移表が指定できる．'$$'以降に構文情報と意味処理を記述する．述語 in で前方形態素までに構成されたＰＳＴ（ＣＩＬ部分項[5]）を受け取り，共通処理を加えた上で後方接続形態素の種類による固有処理を行い述語 send で後方にＰＳＴを送る．ここで send の代りに call_sax と書けば必要なデータがＳＡＸへ送られる．

各処理部はすべてＣＩＬプログラムで記述する．

```
begin(カテゴリ名).
表層 ::   遷移表名a([状態名,..]),
    && [ 遷移表名b([状態名,..]),.. ],
    $$ [ [in(I),共通処理],
         遷移表名b(状態名,[固有処理,send(0)]),
           .... ].
....
end(カテゴリ名).
```

図２　形態素辞書記述形式

今回の開発ツールでは，'$$'以前の連接規則のデバッグ機能のみで，意味構成部のデバッグはサポートされていない．意味構成部のデバッグ機能は，今回作成したツールの機能を拡張する形で実現する予定である．

４．開発環境の構成

図３にＬＡＸの開発環境の構成を示す．マスター辞書[6]から必要な情報を取り出して形態素辞書の元を作り出す辞書トランスレータ，形態素辞書の内容を修正，追加するためのＬＡＸ辞書エディタ，解析プログラムを生成するジェネレータ，形態素解析の結果出力されるデータ構造をインスペクトして連接チェックの再現や構文情報の構成途中のデータなど詳しい内部データを表示するＬＡＸインスペクタなどから構成される．本辞書エディタを用いた編集は直接，辞書記述に対して行うのではなく，辞書記述を変換した中間データファイル群（オブジェクトの辞書クラスに対応）に対して行われる．これは，辞書内容の局所的な変

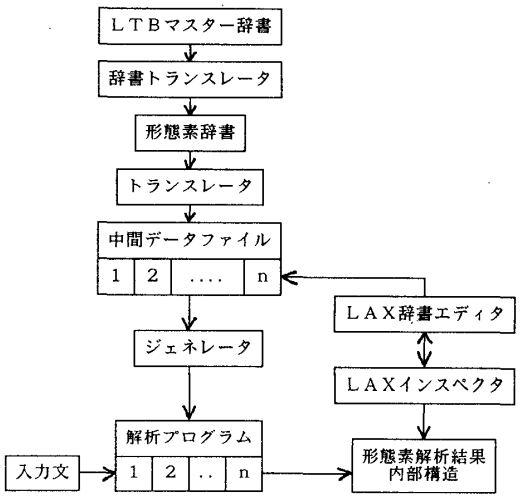

図２　ＬＡＸシステムの構成

Environment of morphological analyzing system LTB-LAX
Yukihiro KUBO　*Masataka YOSHIZUMI　Hiroshi SANO
Kouji AKASAKA　Ryoichi SUGIMURA　ICOT　*JIPDEC

更による連接テストの再試行の時間を最小限にするためである．以下インスペクタと辞書エディタの概要を説明する．

《ＬＡＸインスペクタ》

インスペクタは二つのモードを持つ．一つは解析木表示モードで，このモードでは入力文を形態素解析すると，その解析時間と解析木数ならびに一番目の解釈が表示される．ここでユーザは解析結果のオルタナティブを自由に見ることができる．解析結果は形態素の並びが文節毎に改行して表示され，それぞれ解析に成功した形態素のカテゴリ，状態遷移表名，接続素性名が分かるようになっている．このとき解析結果を一つ一つ指定して見るのではなく一度に見易く表示するプリティプリント機能も用意されている．ユーザは先ずこのモードにより，自分の作成した文法が正しく動作するかどうかが確かめられる．

もう一つのモードが解のインスペクトモードである．このモードへの移行は先の解析木表示モードにおいて，入力文の任意の文字をマウスクリックすることにより行われる．それによりインスペクタは，クリックされた文字から始まる形態素の候補を後方形態素欄に，その文字に連接しうる形態素の候補を前方形態素欄にそれぞれ表示する．ここで表示された任意の形態素を選択することでその形態素に連接し得る形態素とその辞書記述を知ることができる．この機能により，ユーザは文法記述の誤りをいち早く知ることができる．このとき，文法に誤りの見付かった形態素をマウスクリックすることにより辞書エディタをスムーズに起動し記述内容の修正ができる．エディット終了後は自動的にインスペクタに戻るので，修正後の文法による形態素の連接チェックや形態素解析がすぐに行える．インスペクト中のウインドウの例を図４に示す．

《辞書エディタ》

辞書エディタは単独で起動される場合と先のインスペクタから呼ばれる場合とがある．インスペクタから呼ばれる場合にはカテゴリと表層が指定されているので，その形態素の辞書記述がエディットウインドウに表示され，すぐにエディット出来る状態になる．単独で起動した場合には，辞書エントリを検索してから追加変更を加える場合と，まったく新しいエントリをエディットウインドウに入力し追加をする場合がある．新しいエントリの追加の場合，名詞などの様に表層のみが異なり連接規則が同一の形態素の登録を頻繁に行うことが考えられるが，その際は一度元にな

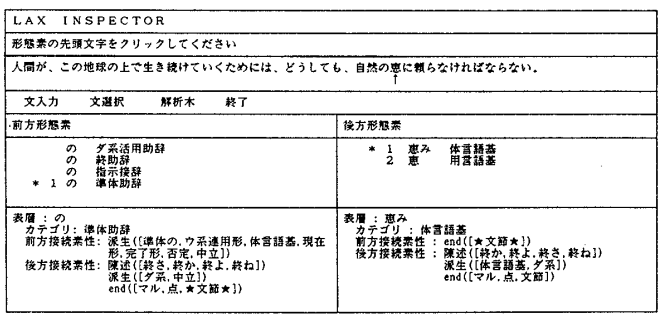

図４　ＬＡＸインスペクタ

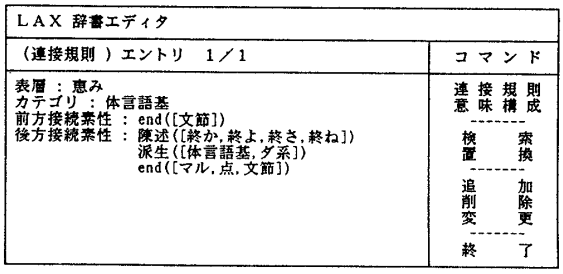

図５　ＬＡＸ辞書エディタ

る連接規則を作成するかあるいは検索により呼出して，その形態素の表層を替えたものを次々に追加できる．追加や変更を受けた形態素の辞書記述は，対応する中間データファイルに書き出され，オブジェクトプログラムの生成まで行われる．これにより辞書エディタの終了後，直ちにインスペクタによる形態素解析のテストが行える．ＬＡＸ辞書エディタのウインドウを図５に示す．

５．おわりに

汎用日本語処理系ＬＴＢにおける形態素解析システムＬＡＸの文法開発環境について述べた．ＬＡＸインスペクタとＬＡＸ辞書エディタを連携して使用することにより形態素文法辞書の開発を効率的に行うことができる．このＬＡＸシステムでは様々なレベルの形態素解析プログラムを作成できる．すなわち，① 形態素切りのみ ② 形態素切りと意味構成（ＳＡＸは未使用）③ 形態素切りと意味構成とＳＡＸコール，という具合にユーザのシステム構成に合った形態素解析プログラムを開発できる．また形態素文法に関しても，① ＬＴＢ標準文法を使用し自立語の追加のみ行う ② ＬＴＢ標準文法仕様を拡張する ③ まったく新しい形態素文法を作成する，などの使用法が考えられる．

今後は，文節意味構成部のデバッグツールの作成など，開発環境の整備を進めていくとともにＬＴＢ標準形態素文法の開発に力を注ぎたい．

【参考文献】

[1] 佐野,赤坂,久保,杉村：語構成に基づく形態素解析，情報処理学会第36回全国大会,1988.

[2] 杉村,赤坂,久保,松本,佐野：論理型形態素解析ＬＡＸ，Proceedings of the Logic Programming Conference '88,pp213-222,1988.

[3] 松本,杉村：レイヤードストリームを用いた並列プログラミング，Proceedings of the Logic Programming Conference '87,pp223-232,1987.

[4] 山崎,弘田,赤坂：ＬＴＢにおける構文解析システムＳＡＸについて，情報処理学会第37回全国大会,1988.

[5] ＣＩＬ言語マニュアル（第３版２刷）ICOT-TM-242,1988年4月

[6] 瀧塚,杉村,田中,赤坂,佐野,重永:ＬＴＢマスター辞書の構成，ソフトウェア科学会，論理と自然言語研究会,1988年12月.