日本語学習支援システムにおける作文診断処理系の試作

掛川 淳一　神田 久幸　藤岡 英太郎　伊丹 誠　伊藤 紘二

東京理科大学基礎工学部電子応用工学科

{kakegawa,kanda,eitaro76,itami,itoh}@itlb.te.noda.sut.ac.jp

1　まえがき

第二言語教育としての日本語教育（以下単に日本語教育という）の主な需要は，日本に在住する人たちが，仕事の必要から，また日常生活上の必要からコミュニケーションをとるため，あるいは，日本以外に在住の人たちが，日本企業の現地事業所に有利に採用されるためや，日本からの観光旅行者の案内を仕事とするためなどであるが，日本企業の海外進出が盛んになるにつれて，後者の需要が大きく伸びている．

言語教育の現場では，文法に偏った教育への反省から，コミュニカティブアプローチに代表される手法が取られている．この手法においては，多様な具体的な状況に対応できる柔軟な言語使用能力を学習者が獲得することを目標とし，場面設定を学習者に与え，そこでの表現の違いの比較を通じて学習する．文法や文型の教育は，これまでに取られていたようなそれだけで独立したものではなく，状況と機能意図に即した言語使用のための言語用法を学習させることになる．

そこでこれまで我々は，このような教育方針に準拠した日本語学習支援システムの作成について検討を重ねてきた [4] [5]．

その中で，LTAG（Lexicalized Tree Adjoining Grammar）を日本語に合わせて定式化し，それに基づいて，学習者の誤りに対する頑強さを持つ診断処理系 [6] [7] [8] を試作したので報告する．

2　日本語の特徴

誤り診断の手法について触れる前に，日本語教育から見た日本語の規則や，特徴に目を向けてみる．

日本語の特徴を以下に挙げる．

1. 文頭から述語に至るまでの語順の自由度が高い．
2. 係り関係は自立語へ係る句がその自立語の前に並ぶ．
3. 格関係の動詞への依存度が大きい．
4. 話者の情報に対する態度を表す文要素であるモダリティ語や文の対話中の機能を表示する表現が，主な述語よりも後部にくる．
5. モダリティ表現に対する副詞の呼応要素が存在する
6. 複合動詞は，その複合動詞の本動詞と必須格が異なることがある．
7. 主題提示部と叙述部からなる提題構文が良く使われる．

3　診断の前提と目標

我々の構想する作文の診断については，

- 言語用法的な誤りの指摘
- 入力文が別の解釈をされる危険がある場合の指摘
- 設定状況においての（不）適切さの指摘

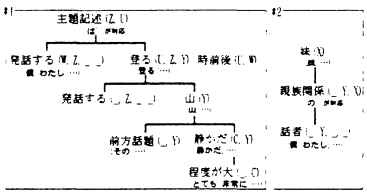
を行うことを最終目標とする．

そして，誤り診断処理系は，具体的な場面設定での前後関係が与えられた穴埋め作文における誤り診断を行なう．

解析には以下の 3 つを制約として用いる．

- 場面の状況（例えば共感度，文脈）
- 正解の意味的係り関係表現
- 意味的係り関係表現の各要素に対応する語彙の候補（学習者が選択可能な語彙リスト）

図 1に正解の意味的係り関係表現と対応する語彙の例を示す．ここで，正解の意味的係り関係表現は，多分木の構造となっており，各ノードに置かれた意味表現（多くの場合自立語，一部機能語が対応）の意味を限定する（多くの場合，係る語の）意味表現を子ノードに置く形をとる．



正解の意味的係り関係表現に対応する入力文の正解例

‡1　僕はそのとても静かな山に登った．

‡2　僕の妹

図 1: 正解の意味的係り関係表現と対応する語彙の例

我々が現在までに診断を試みている誤りは次のものである．

e-1. 係りの不足　正解の意味的係り関係表現が，係ることを要求する意味に相当する語が，係りの位置にない場合

e-2. 係りの障害　正解の意味的係り関係表現が要求する係りを妨げる語がある

e-3. 交差係り

e-4. 意図しない係り　正解の意味的係り関係表現と異なる係り関係

e-5. 接続辞の誤り　助詞の間違いや不足

e-6. 活用の誤り　動詞や形容詞の活用の間違い

e-7. 状況依存の表現の誤り　与えられている状況における「コソアド」表現や授受表現の誤り

4 日本語のための LTAG(Lexicalize Tree Adjoining Grammar)

4.1 TAG(Tree Adjoining Grammar) と LTAG

TAG とはペンシルバニア大学の XTAG リサーチグループによって研究報告 [1] されている文法形式である。

TAG の木のもとになる elementary な木構造は木の継ぎ手の部分の形によって，2 種類に分けられる．それは Initial Tree と Auxiliary Tree（図2参照）である．

また，以上の elementary な木に対して図3の 2 つの操作（substitution と adjunction ）を行うことで，木を成長させていく．

さらに LTAG(Lexicalized TAG) は，このような木をその葉となる辞書項目に書き込んだものを単位として木を成長させていく文法形式である．

4.2 日本語 LTAG の概略

我々はこの文法形式を参考とし，日本語のための，誤り診断処理に適した LTAG を定式化した．

以降では，我々の開発した LTAG について論ずる．

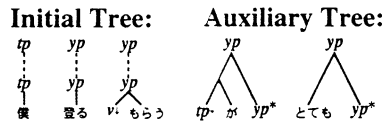


図 2: 日本語の Elementary Tree

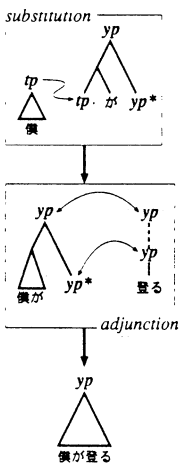


図 3: 操作

図 2に lexicalize された日本語の elementary tree を，図 3に 各操作を例示した．各辞書項目（Lexical item）にはそれを葉とする可能な木構造が書き込まれている．図 2で用いられているノードのラベル tp，yp とは，それぞれ日本語における体言句，用言句を表し，また，継ぎ手としての役割を持たないノード（例えば，pp（：後置詞句）等や，既に充足してしまった継ぎ手）は省略される．

また，木の継ぎ手の役割を果たすノードには素性情報が記述され，例えば，yp の継ぎ手としての形式は

yp (< 主辞変数 >, < 活用形 >, < 格情報 >,
< 未使用格情報 > < 意味素性制約 >,
< 意味的かかり関係表現 >).

である．

また，辞書の形式は木のタイプを表す述語表現で表され，例として図 2で示されている「登る」に関しては以下の通りである．

```
auto('登る',
 yp(X,'終止',[[Z,['人'(Z)],'が'],[Y,['場所'(Y)],'に']],
   SlashCase,['行為'(X)],['登る'(X,Z,Y)|SemXMod])
 yp(X,'終止',_,[[Z,['人'(Z)],'が'],[Y,['場所'(Y)],'に']],
   ['行為'(X)],['登る'(X,Z,Y)])
).
```

述語名 "auto" は自立語（名詞，動詞，形容詞）の有する木のタイプを表し，この他にも，前置語（副詞，指示詞）タイプ "prio"，後置語（モダリティ語，発話行為の助詞）タイプ "post"，接続辞（格助詞，接続助詞）タイプ "link"，判定詞のタイプ "decision"，複合用言のための補助用言のタイプ "appe" が存在する．

LTAG においては辞書項目どうしの木の操作の際に，前述した素性情報は，継ぎ手どうしのユニフィケーションによりその操作が行われたノードの親ノードへと伝播する．この特徴は，文の解析と共に生成においても便利であり，このことが解析と生成を並行して行う必要がある診断処理に LTAG を採用した理由である．

4.3 スタックの利用と SAT と SIT

我々は，句が必ず前から語に係り，句を形成するという日本語の特徴を利用し，LTAG の辞書項目どうしのユニフィケーションで句を作るという特徴以外は，作られた句をスタックに保持し，係ることのできる語を待つという通常の方法 [2] を利用する．

学習者の入力文には一般に誤りが含まれているので，診断の対象となる句について，それが係ることのできる相手（自立語）に合わせた形で誤りを訂正するという作業を要する．このため，仮に形成された句を，スタックに保持し，係ることのできる句を待つという処理形式が不可欠である．

この診断処理系を構築する上で重要な SAT（Saturated Auxiliary Tree）と SIT（Saturated Initial Tree）について定義する．

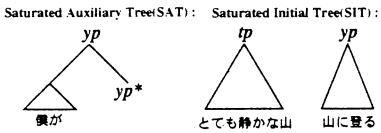


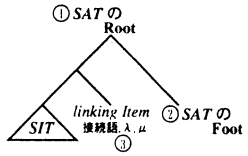
図 4: SAT と SIT の例

SIT とは root 以外の継ぎ手がすべて埋められた木であり，そのデータ形式は，

sit (< 主辞の表層 >, < root の継ぎ手形式 >,
< 主辞を修飾する語の表層 >,
< 誤り情報・誤り訂正情報 >).

とする．自立語の木を出発点に，これに係る句（実は SAT）をスタックから下ろして adjoin し尽くすことにより，SIT が構成される．

SAT とは，root 以外の，foot でないすべての継ぎ手が充足した Auxiliary tree のことであり，SAT は SIT 構成後，先読みにより見出された接続辞を付加することによって作られ，以降の語に係るべくスタックに入り待機する．SAT のデータ形式は，図 5のようにとる．



■■t(<③>,<①の継ぎ手形式>,<②の継ぎ手形式>,<SIT 形式>).

図 5: SAT のデータ構造

なお，SIT の < 誤り情報・誤り訂正情報 > 引数は，以降において述べる誤り診断処理系において，途中の診断結果を記録するためのものである．

4.4 空辞 λ

用言は，SIT の主辞になることができ，接続辞をつけずに活用形によって SAT を構成する事ができる．例えば，連用形は用言にかかる SAT，連体形は体言にかかる SAT を構成する．

このような場合でも，上述した SAT の構成法と形式上の統一をとるために，空辞 λ を接続辞として用いる（図 6）．

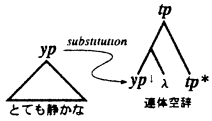


図 6: 連体空辞の例

4.5 Slash カテゴリ

用言の連体形による体言修飾は次のように扱う．日本語の特徴として，前述したように，用言は格を要求するものであるので，連体形をとる用言の格について意味的に該当する句をスタックから下ろせるだけ下ろした結果，充足されない格の情報を Slash カテゴリとして，その主辞の意味素性とこの用言句がかかる体言の意味素性の一致を見，合格すれば，連体形による体言修飾と判断する．

4.6 複合用言

固定した組合せの複合用言については 1 つの語として語彙項目に登録するのが適当であるが，多様な組合せで出現する複合用言については，解析・診断の際に組合せる扱いが必要になる．例えば授受表現その他の複合動詞の木は図 7のように定義される．

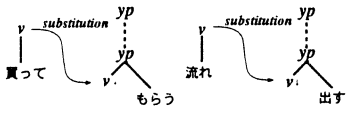


図 7: 複合用言の構造

複合動詞の木を作成するために，複合動詞の本動詞をとることが可能な動詞のて形，もしくは連用形の木は，単純な Initial tree と，図 7の「買って」，「流れ」のような特殊な木の，2 種類の形式をとる．また補助動詞となることが可能な動詞は，単純な Initial tree の形と，「もらう」，「出す」のような形式を取ることとなる．

また，複合動詞については本動詞の格が補助動詞の格支配に従うことが起きる．例えば，

「僕は友だちにこの本を貸してもらいました。」

については，「貸す」に対する「友だちが」・「僕に」が，「もらう」に対する「友だちに」・「僕が（は）」に変わる．

4.7 呼応表現

例えば，以下のような文では，「おそらく」と「だろう」が呼応の関係にある．

「明日は おそらく 雨が降る だろう。」

モダリティ語「だろう」に対して「おそらく」が呼応を許すためには，図 8の構造をとる必要がある．ただし，「おそらく」の辞書には，呼応の相手として “推量” のモダリティ語を要求することを書いておく．

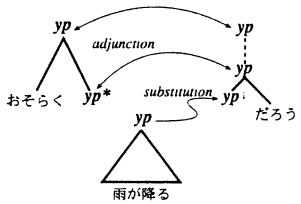


図 8: モダリティ語の構造

5 診断処理系

以下に，我々の誤り診断のアルゴリズムを記す．なお，下記において，(e-j)，j = 1,2,…,7 はに挙げられている該当する誤りの検出・訂正を示している．

1.
   - 学習者の入力した語のリストが空であれば，終了．
   - 空でなければ，学習者の入力文から一語を取り出し，辞書引きを行う．この語を W とする．
     - W が，前置語のように SAT の形式をとるものであれば，main-stack へ push し 1. へ．
     - W が用言であり，かつ意味的係り関係表現を参照し複合用言を形成するいずれかの用言であれば，先読みを行い辞書引きをする．新たに辞書引きされた語を V とする．
       * W V がこの順で当該複合用言の成分であるならば，意味的係り関係表現に照らし合わせて W を（**活用を正して** (e-6)）V に併せ，要求する複合用言が生成される．生成された句を W として 2. へ．
       * V が複合用言を形成する用言でなければ，成分の欠落を，V W の順が正しければ順序の誤りを記録して要求される複合用言を生成する．その語（複合用言の本動詞もしくは補助動詞）は新たに形成された句を W として 2. へ．
     - W が用言であり，意味的係り関係表現を参照し，W が複合用言を形成しない場合でも，先読みを行い，**以降の語に意図せず係ってしまう可能性がある** (e-4) かどうかチェックする．係ってしまう場合は，誤りとして記録する．
       その後，2. へ進む．

- W が用言以外であれば，2. へ進む．

2. W の木を SIT の初期木として，意味的係り関係表現から W への係り情報を取り出して d-list とする．3. へ．

3.
   - main-stack が空であれば，4. へ．
   - main-stack から pop した修飾句の SAT のうち，
     (a)
     - 局所的意味的係り関係表現 d-list に合致するものについて，$\star$ が付いていれば交差係り (e-3) の誤りを記録する．現在見ている SIT の root ノードにおいて，その SIT のもとになった辞書に起源をもつ情報を見ることにより誤り（例えば SAT の**格助詞の誤りや不足** (e-5)，**活用の誤り** (e-6)）を検出して訂正後，SIT に adjoin し，誤りの情報を記録をする．この際，tmp-stack が空でなければ，**係りの障害** (e-2) となる語があったことになるため，さらに誤りとして記録するとともに tmp-stack にある SAT に $\star$ をつける．d-list から adjoin した SAT に該当する要素を除去する．
     - d-list に合致しないものは現在処理している SIT に係ることができるかどうかをチェックし，係ることができれば**意図しない係り** (e-4) の可能性があるので記録する．SAT を tmp-stack へと退避させる．

     (b) 3. へ．

4. d-list を参照し，空になっていなければ，**係るべき語の不足** (e-1) を記録する．
   tmp-stack 内の SAT を tmp-stack が空になるまで，main-stack へと push する．
   先読みし，辞書引きされた語を U とする．
   - U がモダリティ語であれば，現在処理している SIT の主辞が動詞である場合，意味的係り関係表現と SIT の情報から新たなモダリティ句の SIT を構成しうるかチェックを行う．同時に**意図しない係り** (e-4) の可能性も調べる．呼応を見るために 3. へ．
   - U が接続辞であれば，SIT に付加して SAT を作り，スタックに積む．1. へ．
   - さもなければ，
     - 現在処理している SIT の主辞の表層が連体形（連用形）であり，さらに先読みされた語が体言（用言）であれば，空辞 λ を接続し SAT を形成し，スタックへ積み，1. へ．
     - さもなければ，未定辞 μ をこの SIT に付加して SAT を作り，スタックに積む．1. へ．

先読みによって接続辞が見出されない場合，これが接続辞の欠落であるか否か，否としても，連用，連体等の活用形が正しいか，否かが不明であるときには，未定辞 μ を仮接続して SAT を作りスタックに積む．スタックから取り出され，3. で処理されるときに正解の意味的係り関係表現に照らして，係るべき語に相応しい形に修正される．

上記の診断処理 (3.(1) (e-5)) において，格の不足，格助詞の不足，格助詞の誤り等は形成中の SIT の root に登ってきている未充足格の情報と SAT の格情報との照合によって検出できる．

また，活用形の誤りの検出・訂正に関しては，意味的係り関係表現を参照し係ることが判明していれば，tp，yp の情報より相応しい活用形（連体形／連用形）に訂正ができる．

なお，(e-7) については [4] で試みた方法を組み入れる予定である．

6 まとめと今後の課題

以上のような処理系を用いることで，誤りが含まれるような学習者の入力文に対し，誤りの検出された時点で解析を終了するのではなく，意味的係り関係表現に照らし合わせ，最後まで診断を行うことで，学習者の言語用法的知識を測ることが可能となるものと考えられる．

また，現在の処理系で用いている辞書は，試作のため，人手で作成したものであり語彙数は限られる。入手可能な電子化された辞書を変換して利用することを検討している。

主題提示や取立ての「は」に関しての辞書項目の検討を行うことが必要である．

なお，本研究については，文部省科学研究費補助金 09680303 による支援を受けた．

参考文献

[1] The XTAG Research Group(1995): "A Lexicalized Tree Adjoining Grammar for English", University of Pennsylvania, IRCS Report 95-03, March 1995.

[2] 長尾 真(1996): "自然言語処理", 岩波書店.

[3] 田中 穂積："自然言語解析の基礎", 産業図書.

[4] 劉軼, 榎本圭孝, 加藤伸隆, 馬目知徳, 伊丹誠, 伊藤紘二："状況と機能に応じた日本語表現の学習を支援するシステム", 電子情報通信学会論文誌, Vol.J80-D-II, No4 (1997.4)

[5] Nobutaka Kato, Yi Liu, Tomonori Manome, Hisayuki Kanda, Makoto Itami, Kohji Itoh: "Use of Situation-Functional Indices for Diagnosis and Dialogue Datebase Retrieval in a Learning Environment for Japanese as Second Language", Proceedings of AIED '97, pp.247-254(1997).

[6] 加藤伸隆, 神田久幸, 馬目知徳, 伊丹誠, 伊藤紘二："日本語学習支援のための LTAG による文の生成と診断について", 言語処理学会第 4 回年次大会発表論文集, pp.658-661 (1998).

[7] 馬目知徳, 神田久幸, 掛川淳一, 長澤直, 伊丹誠, 伊藤紘二："日本語学習支援における診断のために日本語処理系について", 情報処理研究会報告 99-NL-129, pp.95-100(1999).

[8] 神田久幸, 馬目知徳, 掛川淳一, 長澤直, 伊丹誠, 伊藤紘二："日本語学習支援のための診断処理について", 言語処理学会第 5 回年次大会発表論文集, pp.104-107(1999).

[9] 掛川淳一, 馬目知徳, 神田久幸, 長澤直, 伊丹誠, 伊藤紘二："日本語学習支援のための診断処理系の試作について", 人工知能学会全国大会（第 13 回）論文集, pp.77-80(1999).