多言語資源作成のための文構造タグ付加支援 FLASH アプリケーションの開発

鈴木慎吾[†]　山崎直樹[††]　堀一成[†]
[†]大阪大学 大学教育実践センター　[††]関西大学 外国語教育研究機構

1. はじめに

本稿では、言語コーパスに文構造のアノテーションを施すために現在作成中のアプリケーションについて、その概略を述べる。

今回の発表者 3 名を含むチームは、もともと大阪外国語大学に所属していたメンバーが中心となり、本学（大学統合により現在は大阪大学の一部となっている）における多言語教育のリソースを元に、多言語からなるコーパスを蓄積し、またそれを活用するための方法を以前より模索してきた。まずはじめの段階としては、プレーンテキストと音声データによる多言語平行コーパスを作成してきたが、これは現在までにある程度の蓄積がなされている [1-4] 。今はこのコーパスを言語研究、あるいはさまざまな言語処理に応用する可能性を高めるために、コーパス所収のそれぞれの文について、その構文情報を GDA [5] によってマークアップすることを企画している段階である。

ところで、ここでのマークアップ作業は人手による作業となり、その作業はそれぞれの言語の専門家に依頼することになる。ここにおいて、実際に作業を行うためのツールが大きな問題となる。一般的なテキストエディタはもちろん、既存の XML 用のタグエディタであっても、一般人にとってこのような作業に不便なく使えるとは言い難い。ましてや、我々のようにマイナー言語をも含めた多言語コーパスを作業対象にしていると（本学外国語学部には専攻語だけでも 25 の言語があり、我々はそれらのうちできるだけ多くの言語を対象にしたいと考えている）、該当言語の文構造が理解できる、我々の周りにいる数少ない作業候補者を確保するためには、できる限り学習負担の少ないツールを開発しておくことがどうしても必要になる。

同様の目的を満たすために作成されたツールにはすでに eBonsai があり、我々の作業にとって大いに参考になっている [6] 。eBonsai は自動解析機能を備えた本格的なツールであるが、我々は特に多言語対応を指向し、なおかつ、マークアップツールとしてのみならず、特にツリー構造の描画部分を工夫し、外国語教育や人文系言語研究の場面で広く活用することができるようなものを目指している。

2. タグ付けの実例

例えば、次の中国語に統語構造情報を付与する場合を考える。

她不小心用锅子把她公公砸死了。

この文は GDA によって以下のようにタグ付けすることができる。

```
<su>
  <np>
    <n>她</n>
  </np>
  <v>
    <adp>
      <ad>不小心</ad>
    </adp>
    <v>
      <vp>
        <v>用</v>
        <np>
          <n>锅子</n>
        </np>
      </vp>
      <vp>
        <v>把</v>
        <np>
```

```
            <np>
              <n>她</n>
            </np>
            <n>公公</n>
          </np>
        </vp>
        <v>
          <vp>
            <v>砸死</v>
          </vp>
          <v>了</v>
        </v>
      </v>
    </v>
  </su>
```

このようなデータを初心者にも簡単に作成できるツールを作ることが今回の初期目標である。

3. 今回開発したソフトウエア

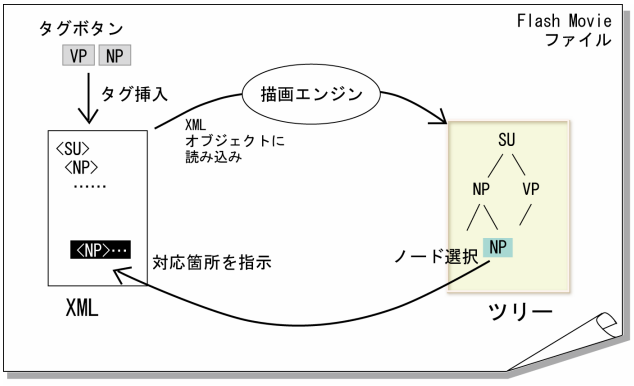
3.1 概要

入力文（平文）について、マウスを使ったGUI 操作によって簡単に構文木を作成する。同時に、そのツリー構造に対応する XML データを出力する。

なお、本ツールは Flash ムービー形式で作成しているので、Flash Player 上で動作する。

3.2 処理の流れ

入力文を ActionScript の XML オブジェクトに読み込んで解析し、ツリーを描画する。ツリーの節点や本文テキストは選択することが可能で、これらを選択するともとの入力文の対応箇所を選択状態にし、タグボタンによってタグを挿入するなどの処理を行うことができる。また、XML データが変更されるとその都度ツリー表示も再描画される（【図 1】）。

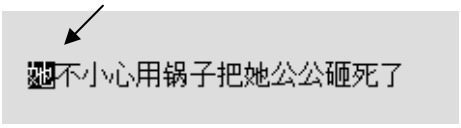


【図 1】処理の流れ

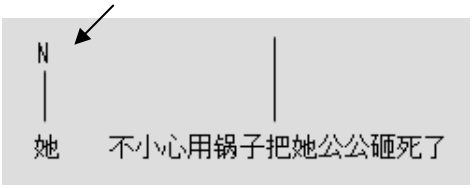
3.3 操作方法

本ツールを使って入力文にマークアップを施す方法は以下の通りである。（全体画面【図 5】において、画面左側を「出入力エリア」、右側を「作業エリア」と呼ぶ。）

1. 出入力エリアに処理したい文（平文）を入力する。（→作業エリアに入力文が描画される）
2. 作業エリアに描画された文につき、タグを付けたい構成素を選択、ハイライトさせた状態【図 2】で、上部に並んでいるタグのボタンを押す。（→親ノードが作成される【図 3】）

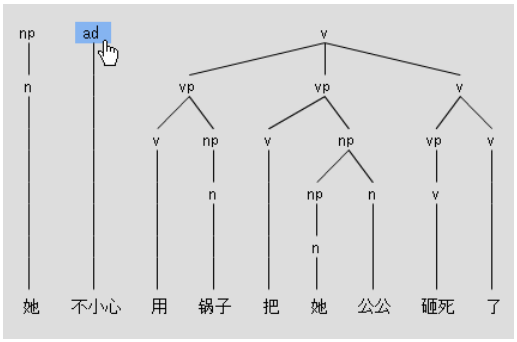


【図 2】構成素をハイライト



【図 3】親ノードが作成された

3. タグのラベルをクリックすることでそのタグノードを選択することができる【図 4】。その状態で上部のタグボタンを押すことによりさらに上位のノードを作成することができる。また、隣り合う兄弟ノードは同時選択することもできる。



【図 4】タグノードを選択

4. 作成されたツリー構造に対応した XML データが出入力エリアに逐次出力される。

3.4 補足説明

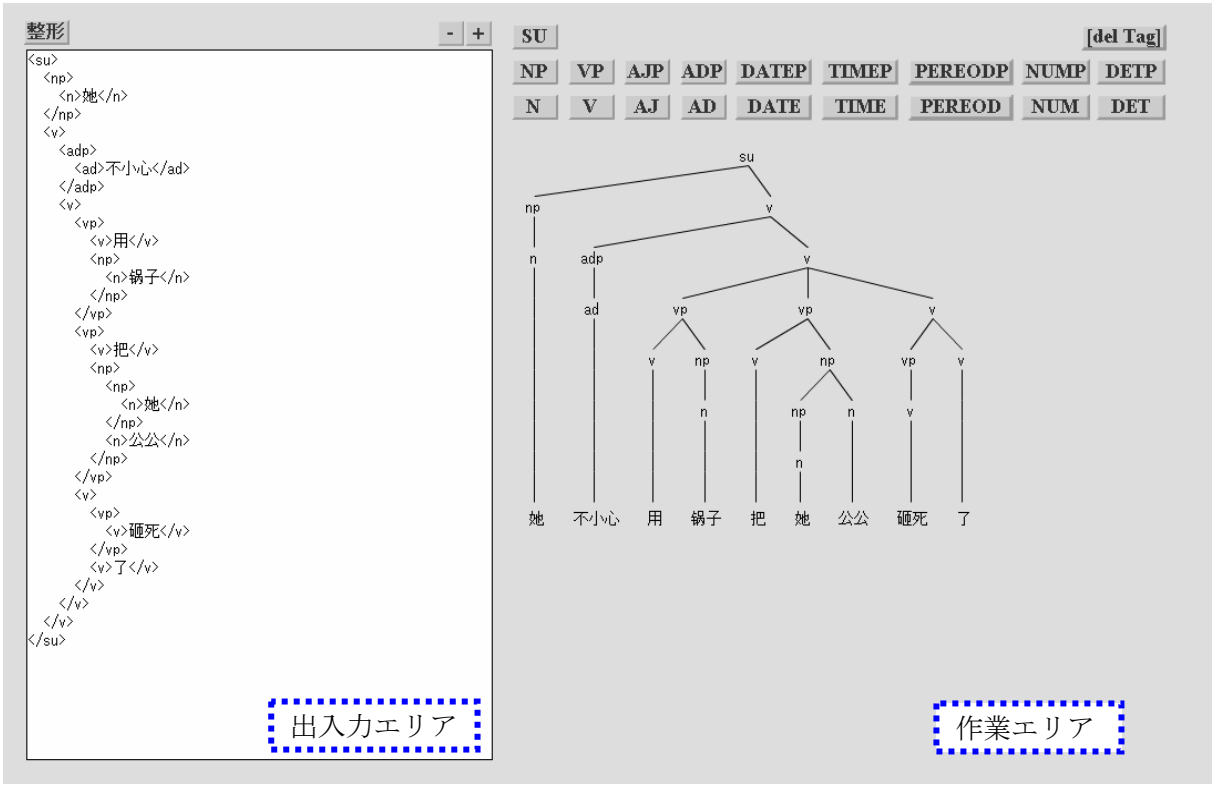
上記の他、付記すべき機能は以下の通り。

- タグ付けに使用するボタン（タグボタン）はタグセットによって切り替えることができる。
- タグのラベルを選択した状態で「del Tag」ボタンを押すと、そのタグが削除される。
- すでに XML でタグ付けされたデータを出入力エリアに入力すると、対応する構文木が作業エリアに描画される。
- （上記に関連して、）出入力エリアで XML データを直接編集することもできる。その際、構文木もリアルタイムに更新される。
- 本ツールは（上で紹介したとおり）XML データをツリー形式で表示する。このツリー表示機能、またツリーの GUI 操作による XML 編集機能は、自然言語の統語構造だけでなく、XML で記述されたデータ一般に関しても広く用いることができる。

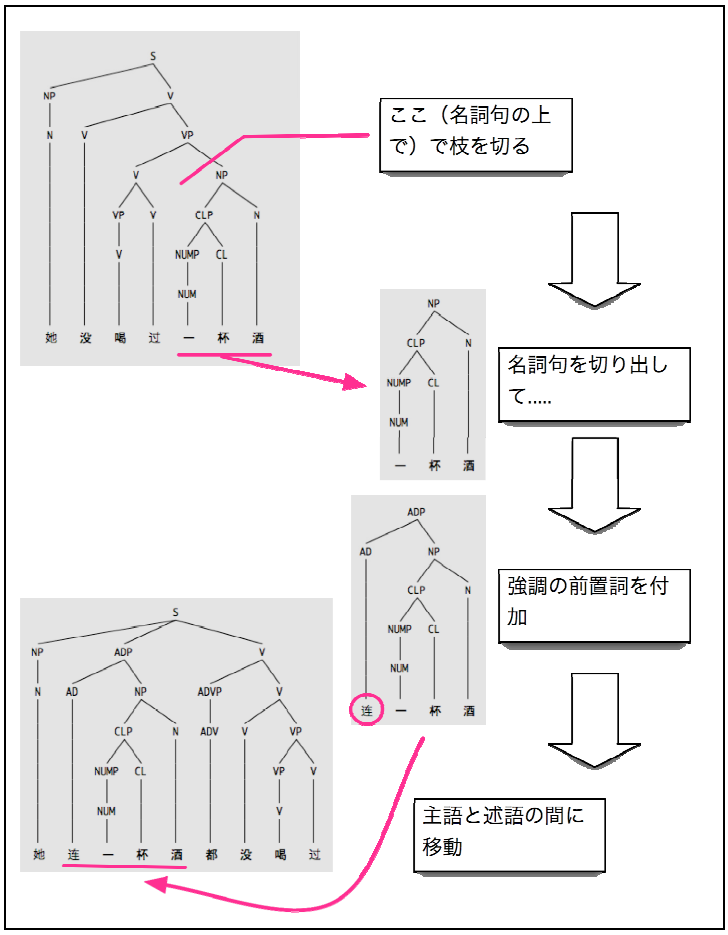
4. 応用と課題

4.1 外国語教育の現場で

外国語教育の現場では、文法構造に対応した構文解析木を学生に示せたら教学の効果が上がるだろうと思われる場面が多々ある。問題は、どうやってツリーを「描く」かにあっ



【図 5】全体画面



【図 6】教材例（中国語の強調構文の生成を説明）

た。本ツールを使用することで簡単に構文解析木を描くことができる。【図 6】にツリー画像を教材に貼り込んで活用した例を示す。

なお、本ツールは Flash ムービー形式であるので、そのまま Web ツールとして外国語教育に従事する同業者に広く提供することが可能である。

4.2 課題

本ツールはまだごく簡単な機能を備えているに過ぎない。今後の課題としては、主に以下の点が挙げられる。

- 属性操作。
- アラビア文字など、右から左へ書く文字への対応。
- Undo 機能。
- さらに柔軟なツリー操作。例えばノードの削除、移動など。
- 依存構造木への変換機能。
- 構文解析の半自動化。

また、実際にコーパスを用いて作業していくうちに改善すべき点も出てくると思われる。

謝辞

本研究は、科学研究費補助金　基盤研究（B）　課題番号: 19300047『LCTL を含む多言語平行マルチメディア資源の構築と構造化方式の研究』（研究代表者: 堀 一成）の補助を受け推進したものである。

参考文献

[1] 堀一成, 石島悌「PostgreSQL による多言語単語データベースの構築」, 情報処理学会第 62 回全国大会講演論文集(2), 2001.3, pp. 297-298.

[2] 堀一成, 前田彩, 石島悌「PostgreSQL と JSP を用いた多言語データベース検索アプリケーションの構築」, FIT2002 一般講演論文集(2), 2002.9, pp. 95-96.

[3] 堀一成「大阪外国語大学の言語資源を用いた言語 e-learning の構想」, 言語処理学会第 10 回年次大会ワークショップ「e-Learning における自然言語処理」論文集, 2004.3, pp. 13-16.

[4] 堀一成, 山崎直樹, 竹原新, 小島一秀「多言語平行マルチメディア言語資源の構築」, 言語処理学会第 13 回年次大会発表論文集, 2007.3, pp. 768-771.

[5] 「大域文書修飾 Global Document Annotation (GDA)」http://i-content.org/gda/

[6] 野口正樹, 市川宙, 橋本泰一, 徳永健伸「構文木付きコーパス作成支援統合環境 eBonsai の新しいインターフェース」, 言語処理学会第 12 回年次大会発表論文集, 2006.3, pp. 751-754.