多言語資源作成のための統語属性付与支援 FLASH アプリケーションの開発

鈴木慎吾[†]　山崎直樹[††]　堀一成[†††]
[†]京都産業大学 外国語学部　[††]関西大学 外国語学部　[†††]大阪大学 大学教育実践センター

1. はじめに

本稿では、言語コーパスに言語情報のアノテーションを施すために開発中のアプリケーションについて述べる。

大阪外国語大学（現大阪大学外国語学部）で発足した多言語処理プロジェクトでは、以前より多言語による平行コーパスの蓄積を行ってきた [1]。また同時に、これらのコーパスを言語処理の分野に応用することを目的に、構文を中心とした言語学的情報をコーパスに埋め込むマークアップの研究も継続して行っている。

この、言語学的情報のマークアップを実際に行おうとする時には、「何を」「どのような形式で」「どうやって」行うのかが問題となる。

まず、「何を」埋め込むかということについて、われわれは特に言語研究および言語教育の立場から、言語間の多様性と共通性をより端的に映し出す情報を優先する方針を掲げている。これに関する基礎研究は、山崎 2009 を参照のこと [2]。

次に「形式」であるが、これには本研究の目的と合致し、汎用性も期待される GDA [3] を本命に据えて検討している。ただし、われわれが対象とするコーパスには様々な言語種が含まれるため、それらに適応できるようタグをカスタマイズすることも考えている [4]。

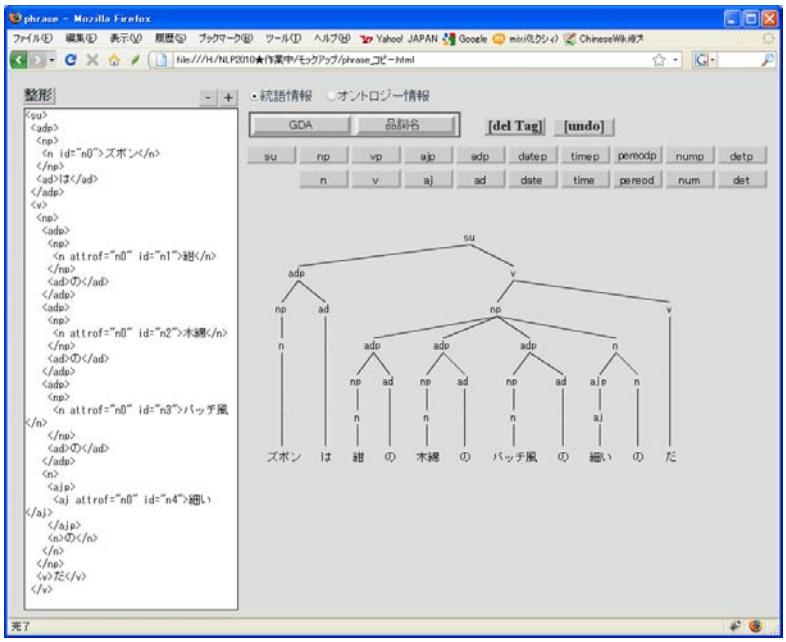
最後に、実際に「どうやって」タグを埋め込むかであるが、データ構築作業において最も重要なことは作業者が直感的に操作できるかという点である。この点を踏まえ、われわれはこれまでに木構造作成ツールおよびオントロジー情報付与ツールを試作し、公開してきた [5, 6]。これらはともに画面上のポインタによる GUI 操作によって言語学的情報を付与するための FLASH アプリケーションである。本論文は主にこれらのアプリケーションの新たな展開と統合について述べるものである。

2. アプリケーションの新たな展開と統合

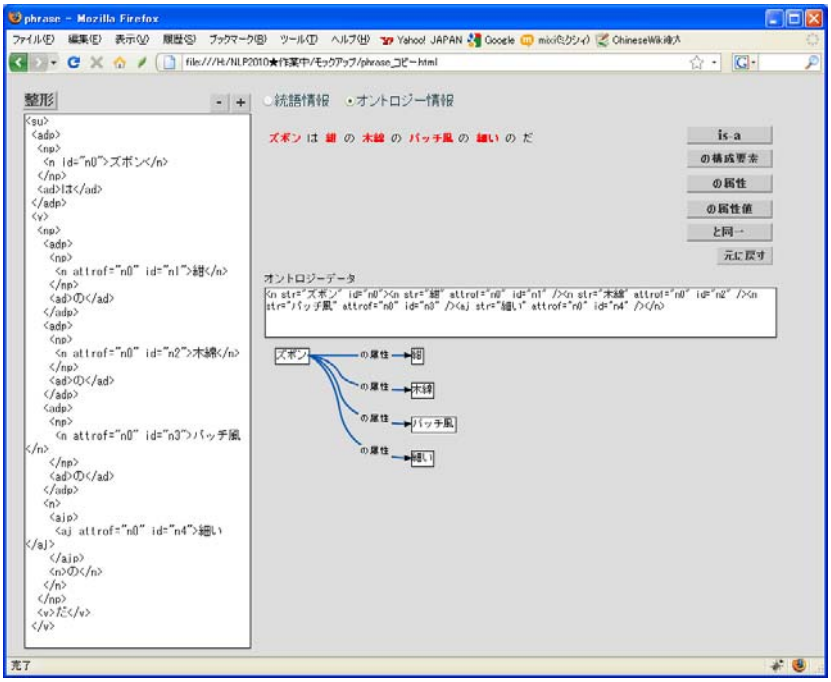
今回紹介するツールはすでに作成した 2 種のツール（注 [5][6] を参照）をベースとしている。開発環境は FLASH CS3（Action Script のバージョンは 2.0）を使用した。

2.1 木構造作成ツールとオントロジー情報付与ツールの統合

今回は、これまで開発してきた 2 種のツールを統合し、一つの画面で操作ができるようにし



【図 1】木構造編集モード



【図 2】オントロジー情報編集モード

た。それぞれの操作画面はラジオボタンによって切り替えることができる（【図 1】【図 2】）。

構文情報とオントロジー情報はいずれも XML によって元データに埋め込むようになっているが、これらは排他的でないため一つのデータ内に共存させることが可能である。具体的には、構文情報は要素で記述する一方、オントロジー情報は属性で記している（下記ソース参照）。

```
<su>
 <adp>
  <np>
   <n id="n0">ズボン</n>
  </np>
  <ad>は</ad>
 </adp>
 <v>
  <np>
   <adp>
    <np>
     <n attrof="n0" id="n1">紺</n>
    </np>
    <ad>の</ad>
   </adp>
   <adp>
    <np>
     <n attrof="n0" id="n2">木綿</n>
    </np>
    <ad>の</ad>
   </adp>
   <adp>
    <np>
     <n attrof="n0" id="n3">パッチ風</n>
    </np>
    <ad>の</ad>
   </adp>
   <n>
    <ajp>
     <aj attrof="n0" id="n4">細い</aj>
    </ajp>
    <n>の</n>
   </n>
  </np>
  <v>だ</v>
 </v>
</su>
```

2.2 XML 属性の編集機能

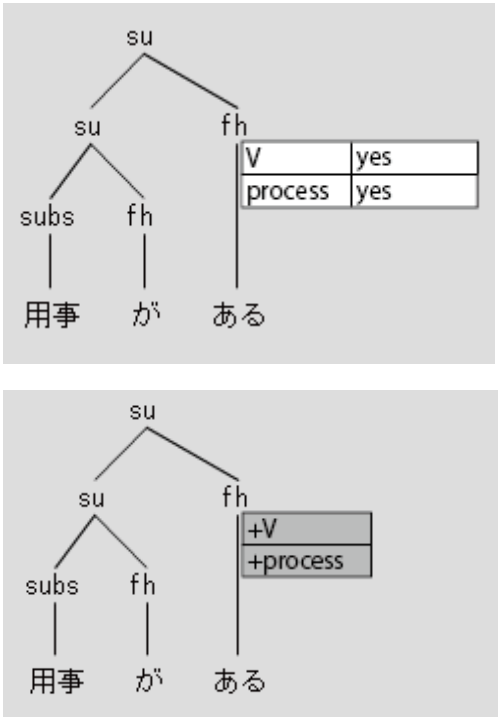
前回までに作成したツールは GUI で属性を編集する機能を搭載していなかった。今回はこれを搭載することにした。

XML の属性は、数が多くなるほどデータが複雑化し、手作業による編集が煩雑になってしまう。この部分を GUI 化することは、実際のデー

タ構築作業において大きなメリットがあると考えられる。

本ツールでは、編集画面に表示されているツリー上で、任意の節点をクリックするとポップアップ画面が出現し、その要素にぶら下がっている属性が表示される。ここで属性値を編集すると、元データにその内容が反映されるようになっている。新しい属性を加えることもできる。

ところで、このポップアップによる属性表示には、編集モードと閲覧モードの二種の表示形式がある。編集モードは XML の属性をそのままに表示するが、閲覧モードにおいては簡単な変換規則を介すことで表示方法を工夫し、言語情報を分かりやすく示すようになっている。これは主に言語学者の需要を満たすために搭載された仕様である（【図 3】）。



【図 3】（上）編集モード
（下）閲覧モード

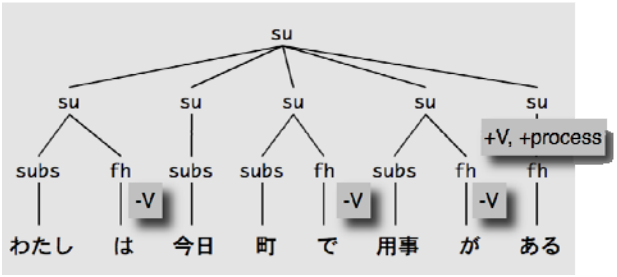
3. タグセットの拡張

前節では開発中のツールの改良点について述べた。本節では本ツールを使用して構築するデータに使うタグセットの特徴について、特に XML の属性の役割に重点を置いて述べる。

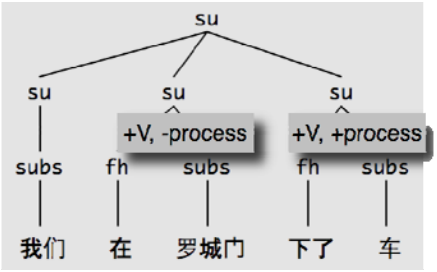
3.1 使用するタグセットの特徴

われわれは、XML で言語学的構造を記述するには、範疇名は最小限に抑え、一方で、その他の細かな情報は属性で記述するのが有利であると考える。その理由を以下で説明する。

まず、現在、仕様を策定しているタグセットは、次の特徴を持つ。



【図 4】日本語木構造の例



【図 5】中国語木構造の例

1. 「文」という上位節点を持たない。いわゆる「文」は、単数のあるいは複数の su (Speech Unit、発話単位) から構成される。
2. su は、subs（Substantive、実詞）のみか、あるいは、subs および fh (Functional Head、機能的主要部）からなる。
3. 日本語の述語動詞と後置詞はともに<fh>である。これに [±v]（動詞性の有無）という属性値を与え、両者を区別する。【図 4】の「は」「で」「が」「ある」を参照。
4. 中国語の述語動詞と前置詞はともに<fh>である。中国語の前置詞はほとんどが動詞由来のものであり、現在、文法化の途上にある。ともに [+v]（動詞性あり）という属性値を持つ。しかし、前置詞として使われる動詞由来の語は、ふつうアスペクト標識を持たない。よって、前置詞的に使われている動詞性の語と述語として使われている語を、[±process]（過程性

	日本語の述語動詞	中国語の述語動詞	中国語の前置詞	日本語の後置詞
範疇名	fh	fh	fh	fh
属性[v]	+	+	+	-
属性[process]	+	+	-	-

【表1】日中両言語における動詞～前置詞のアノテーション（案）

の有無）という属性値で区別する。【図 5】の“在”“下了”を参照（“了”がアスペクト標識）。

日本語の後置詞も中国語の前置詞も、ともに、動詞の取る項が動詞に対してどのような意味役割を持つかを表示する役目を果たす（おおざっぱにいうと、「格関係を表示している」となる）。ここで、中国語の前置詞に、日本語の後置詞と同じ範疇名――例えば、<adp>（Adpositional、側置詞）――を与えてしまうと、中国語内部における範疇間の連続性（動詞～前置詞）が、離散的になってしまい、中国語という言語の特徴の 1 つが見えなくなる。

しかし、この中国語の前置詞に、動詞と同じ範疇名を与えると、こんどは、この動詞由来の前置詞が、（日本語の助詞がそうであるように）格表示の機能を担っているという、通言語的に見て重要な情報（多言語平行コーパスにおいては、ある機能が、A という言語ではどのような統語範疇によって担われ、別の B という言語ではどのようになっているか、は極めて重要な情報）が、見えなくなってしまう。

そこで、本研究で提案する「最小限の範疇」「豊富な属性」の利点が明らかになる。【表 1】でわかるように、範疇間の連続性がよく捉えられている。

必要に応じて、この属性をさらに増やすことにより、より多くの差異を記述しつつ、範疇間の連続性の微視的な推移をも記述できるようになることが期待される。

3.2 属性のその他の使い道

このタグセットでは、以下の情報もアノテーションをすることを考えている。

- 照応情報（定・不定、束縛されている／いない.....）
- 形態変化による格表示の情報（主格、対格、斜格.....）
- 動詞に対する意味役割（動作者、受益者.....）

これらの情報をすべて、属性で表現する予定である。

附記

本稿で述べたアプリケーションは以下のサイトにて公開している。「大阪大学　多言語資源研究グループ」http://mldb.cep.osaka-u.ac.jp/

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。

注（参考文献）

[1] 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.

[2] 山崎直樹「多言語平行コーパスのための「言語学的におもしろい 100 の文」」, 『外国語教育研究』（関西大学）, 2009.3, pp.111-125.

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

[4] 山崎直樹「多言語平行コーパスのための言語横断的な構造記述」, 2010 年中日理論言語学研究国際フォーラム（於同志社大学, 2009.07.26）.

[5] 鈴木慎吾, 山崎直樹, 堀一成「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」, 言語処理学会第 14 回年次大会発表論文集, 2008.3, pp.265-268.

[6] 鈴木慎吾, 山崎直樹, 堀一成「テキストコーパスにオントロジー的知識を付与するための FLASH アプリケーションの開発」, 言語処理学会第 15 回年次大会発表論文集, 2009.3, pp.172-175.