RN-002

多言語資源作成のための統語・オントロジー情報を付与するアプリケーションの開発

Development of an Application to Add Syntactical and Ontological Information to Multi-language Resources

鈴木 慎吾† 山崎 直樹†† 堀 一成†††
Shingo SUZUKI Naoki YAMAZAKI Kazunari HORI

1. はじめに

本稿では、「大阪大学多言語資源研究グループ」が開発した、多言語コーパスに言語情報のアノテーション（情報付与）を施すためのアプリケーションについて述べる。

言語研究を専門にする者が実際に XML をベースとするアノテーションを行う際、XML エディタなどを用いて直接 XML データを作成することはなかなか困難である。そのため、その作業の助けとなることを目的として本アプリケーションを開発した。FLASH アプリケーションとすることで Web ブラウザ上から気軽に作業ができる点と、ツリー描画機能により、言語研究者になじみの深い木構造表示を確認しながら作業ができる点が大きな特徴である。また、単純な木構造では表しづらいオントロジー情報も属性として格納することができ、その作業専用の画面も有している。

1.1 背景

「大阪大学多言語資源研究グループ」は、大阪外国語大学（現大阪大学外国語学部）で組織されていた言語資源の構築に関する研究グループで、現在では大阪大学の研究者を中心としてこれまでの研究成果を引き継ぎ、新たにプロジェクトを立ちあげて研究を続けているものである。本グループの目的は、世界中の様々な言語について、言語学者の知見を結集し、言語処理に必要な資源を構築して、応用分野に発展させる基盤を整備することにある。

この目的を果たすため、本グループは組織を「コーパス構築部門」「マークアップ形式策定部門」「アプリケーション開発部門」の三つに分け、それぞれが連携しながら作業を行っている（図 1）。

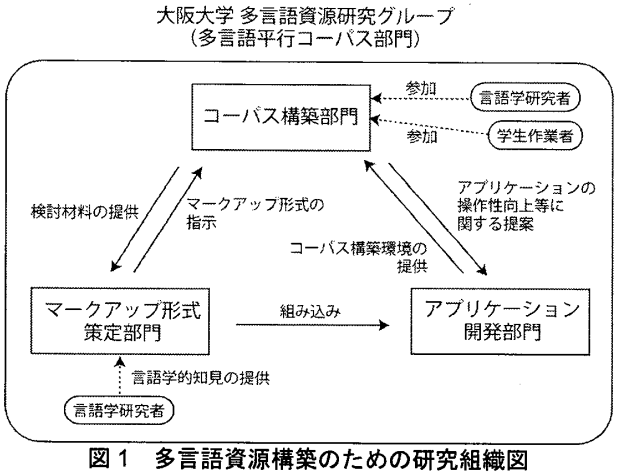

図 1 多言語資源構築のための研究組織図

「コーパス構築部門」は、コーパス構築のための実際的作業を行う部門である。本グループは大阪大学外国語学部を中心とした幅広い人的リソースを活用し、多様な語種によるコーパスを構築することを方針の一つとしており、学部が専攻語として設置している 25 の言語について、まずは基本単語 5000 語と、旅行ダイアログ 1000 文による平行コーパスの構築を目標とし、作業を行っている[1]。

次に「マークアップ形式策定部門」は、上記コーパスに言語学的情報を付与するためのマークアップの形式を検討する部門である。ここではまず、①平行コーパスに記述すべき、言語学的に、あるいは言語処理を行う上で重要な情報とは何かを吟味し[2]、②それらの情報をどのような形式でコーパスに付与するかを検討している。特に②については、目下 GDA [3] をベースに検討を進めているが、これによって多言語コーパスを並行的に扱おうとすると細かな問題が色々と発生することが分かってきたため、我々の目的に合わせるべくタグ定義の修正を検討しているところである。

最後の「アプリケーション開発部門」は、上記「マークアップ形式策定部門」によって策定されたマークアップ形式に基づきながら、実際にコーパス構築に使用するアプリケーションを開発する部門である[4,5,6]。本稿ではここで開発したアプリケーションを紹介する。

1.2 本アプリケーションの特徴

上記「コーパス構築部門」が扱う言語種は、マイナー言語も多く含むものであるため、人手による作業が多い。またそれら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。従って、本ツールは操作方法の習得ができる限り容易であることを目指した。

コーパス向け汎用アノテーションツールとしては、SLAT が知られる[7]。本稿のアプリケーションは言語研究者になじみのある木構造を直接操作するインターフェースを採用しているところに特徴がある。

また、オントロジー構築ツールとしては法造 [8] や Protégé [9] などがよく知られている。それらのツールは純粋にオントロジーを構築することを目的とした機能豊富なアプリケーションである。本稿のアプリケーションはオントロジーを構築するのにテキストコーパスを出発点にしている点と、得られたオントロジー情報をもとのコーパスに埋め込むという点に特色がある。

ところで、ここで開発するツールはコーパス構築のみならず、外国語教育や人文系言語研究といった場面でも広く使える可能性があり、グループの内外からもそのような要望が強い。開発に当たってはそのようなことも強く意識している。

† 京都産業大学, Kyoto Sangyo University
†† 関西大学, Kansai University
††† 大阪大学, Osaka University

2. マークアップデータ形式

本アプリケーションがコーパスに付与する情報は、統語構造情報とオントロジー情報の二種である。例えば、次の文は下に示すような形でマークアップされる。

「彼の車はシートが革張りの赤いスポーツカーだ。」

```
<su>
 <adp>
  <np>
   <adp>
    <np>
     <n>彼</n>
    </np>
    <ad>の</ad>
   </adp>
   <n id="n0">車</n>
  </np>
  <ad>は</ad>
 </adp>
 <vp>
  <np>
   <adp>
    <ajp>
     <adp>
      <np>
       <n partof="n4" id="n1">シート</n>
      </np>
      <ad>が</ad>
     </adp>
     <aj attrof="n1" id="n2">革張り</aj>
    </ajp>
    <ad>の</ad>
   </adp>
   <ajp>
    <aj attrof="n4" id="n3">赤い</aj>
   </ajp>
   <n is_a="n0" id="n4">スポーツカー</n>
  </np>
  <v>だ</v>
 </vp>
</su>
```

2.1 統語構造情報

統語構造の情報は基本的に GDA に準じたマークアップ方式で記述している。ただし、先に述べたように、「マークアップ形式策定部門」の今後の議論によっては修正が加えられる可能性がある。

具体的には、「名詞」「動詞」という大雑把で離散的な範疇を使わず、素性の行列により、範疇間の連続性を、あるいは、異言語間の普遍性と個別性を表現できるようなタグセットを考えている。

例えば、中国語における前置詞（在、把など）はほとんどが動詞由来のものであり、動詞としての特徴を失う文法化の途上にある。ここで、「接置詞」という範疇を立てた場合、中国語の前置詞は日本語の後置詞などとともに「接置詞」に入ることになるが、これでは中国語における動詞と前置詞の関係が分断されてしまい、中国語の品詞間における重要な関係性が記述できない。そこで我々は、範疇の種類は最小限にとどめ、属性の集合によって品詞を表現する方式を検討している。例えば、日本語の動詞・後置詞、中国語の動詞・前置詞の範疇名を全て fh（Functional Head、機能的主要部）とし、属性として動作性の有無（[±V]）と過程性の有無（[±process]）を設定してやると、それぞれは次のように表現される。

日本語の述語動詞：fh [+V] [+process]
中国語の述語動詞：fh [+V] [+process]
中国語の前置詞：　fh [+V] [-process]
日本語の後置詞：　fh [-V] [-process]

このようにすれば中国語の述語動詞と前置詞の連続性をうまく表現することができる。

2.2 オントロジー情報

オントロジーとは本来、現実世界から抽象された「概念」を対象とするものであるが、本稿では「概念」ならぬ「語彙」、それも具体的なテキストに出現する「語彙」を扱う。したがって、本稿が「オントロジー」として構築しようとしているものは、本来の意味における「オントロジー」そのものではなく、個々のオントロジーのインスタンスの関係を記述したもの、つまりオントロジーが現実世界において具現化したところの個別的事象間の関係をとらえたものである。これは本研究がコーパスを出発点としていることによるが、一方で言語形式と具体的事象との関係は言語学の大きなテーマの一つであり、本研究は特にその方面での応用を期待している。また、本稿のような方法は、すでにインフォーマント（現時点での話者）のいない、文献しか資料のない言語（例：各種古典語）について、残された文献からそこに記述されている世界のオントロジーを構築するのにも有効と思われる。

本アプリケーションで扱うオントロジーの意味リンクとしては「～の一種(is-a)」、「～の一部(partOf)」、「～の属性(attrOf)」、「～の属性値(valueOf)」、「～と同じ(equalOf)」の 5 つを用意している。これらを XML の属性を用いて記述する。具体的には、関係を記述する要素の全てに id 属性を付け、概念的に下位の側の要素に関係性を示す属性を記述している。

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んでいる。これらは排他的な記述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。

3. アプリケーションの詳細

前節に示したようなデータの作成を支援する目的で開発したのがここで紹介するアプリケーションである。

本アプリケーションの開発の目的を考えると、インターフェイスは GUI であることが求められる。開発環境は比較的手軽に GUI アプリケーションを開発できる点を考慮し、FLASH CS3 を使用している（Action Script のバージョンは 2.0）。

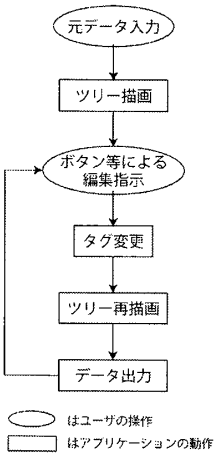

図 2　アプリケーションの動作概略

本アプリケーションの動作の概要は図 2 の通りである。

元データは平文テキストでも XML タグ付きテキストでもよい。タグ付きデータを入力した場合には、直ちに対応するツリーが描画される。平文の場合は、構造情報がないため最初ツリーは表示されない。

ボタンによる編集指示とは、統語情報の編集指示、あるいはオントロジー情報の編集指示である。いずれもマウスによる GUI 操作が可能である。情報が埋め込まれたデータは逐次出力される。

編集画面には統語構造編集画面とオントロジー情報編集画面の二つがある（図 3、図 4）。

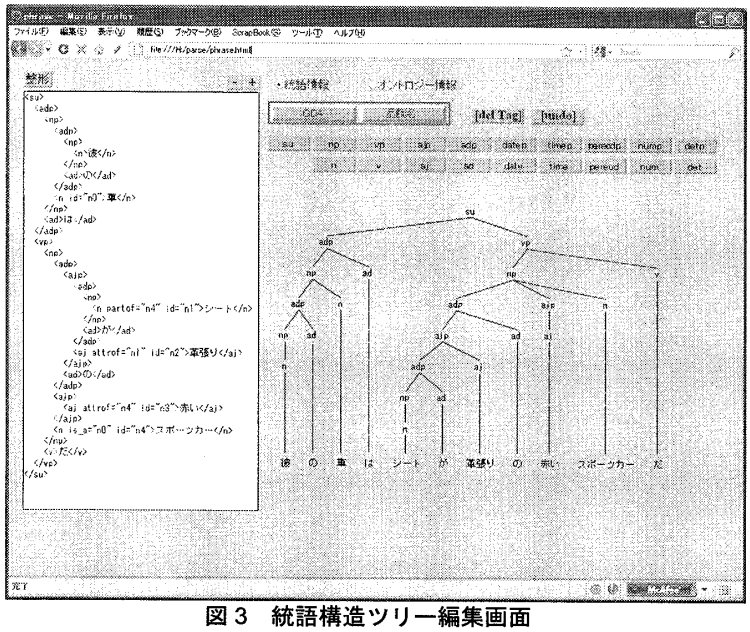

図 3　統語構造ツリー編集画面

3.1 統語構造ツリーの編集

描画されたテキスト、あるいは節点のタグ名を選択し、画面上部にあるタグ名が書かれたボタンを押すと上位ノードが作成される（図 5）。隣り合う節点は同時に選択することができ、その状態でタグ名のボタンを押すとそれらを子ノードとする上位ノードが作成される。この操作を繰り返すことで文全体の構造を埋め込むことができる。

タグ名のボタンは既設のものだけでなくセット単位でカスタマイズが可能である。

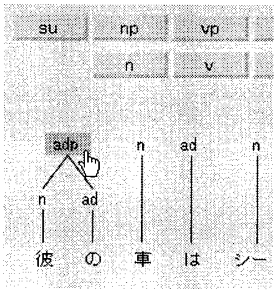

図 5　節点を選択

3.2 属性の表示・編集

ツリー上の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。空欄に新たな属性を付け足すこともできる（図 6）。

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 7）。

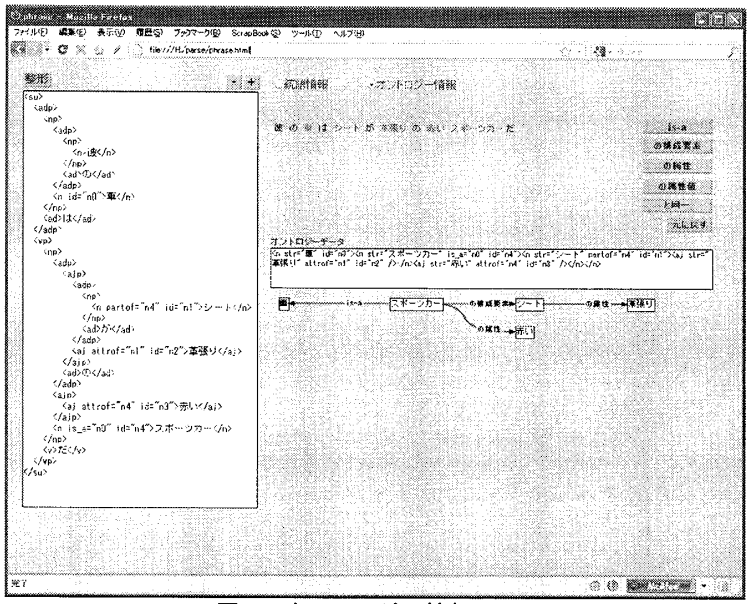

図 4　オントロジー情報画面

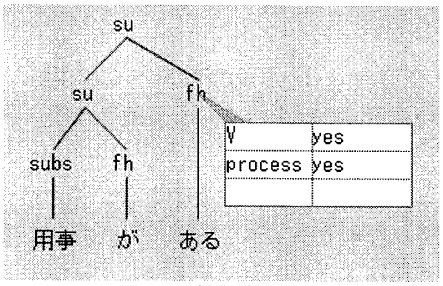

図 6　属性編集モード

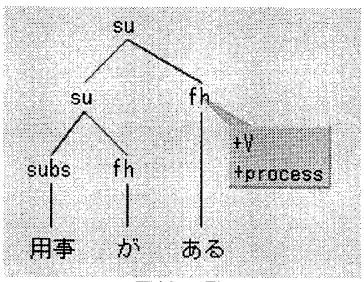

図 7　属性閲覧モード

3.3 オントロジー情報の表示・編集

図 3 の上部にあるラジオボタンをクリックすると画面の右半分がオントロジー編集画面に切り替わる（図 4）。

統語構造の編集画面においてすでにテキストが形態素に切り分けられていれば、オントロジー編集画面に形態素ごとに区切られたテキストが表示される（図 8）。ここで、これらの語彙から任意の二つを選択した状態で、関係が表示されているボタンを押すと、これら二つの語彙を結ぶ意味リンクが作成されるようになっている。

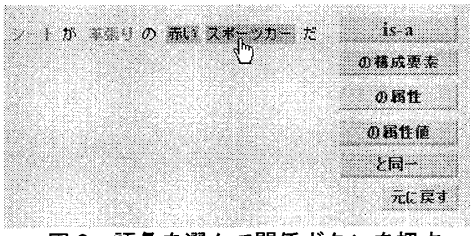

図 8　語彙を選んで関係ボタンを押す

本アプリケーションでは関係ボタンを押した際に、関係の方向をたずねるダイアログを出すようにしている（図 9）。これは、意味リンクを作成する際、作業者がリンクの方向を間違えてしまうことが多いからである。マークアップのミスを減らす上でこのような言葉による確認は有効であることがこれまでの運用実績から確認されている。

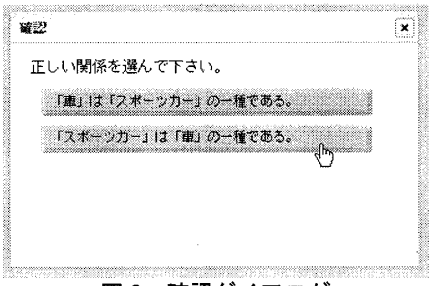

図 9　確認ダイアログ

画面の下部には形成された意味リンクの全体図が表示されるようになっている。

4. アプリケーションの教育への応用

この「所与のテキストから構築したオントロジー」は、外国語学習者の支援ツールとしても使える。1 例を挙げれば、以下のとおりである。

読解対象の長文に出現する語彙を単にピックアップしただけの語彙リストは、高度な内容の読解になるほど、内容把握にはさして寄与しないことはつとに知られている。オントロジーとして提示された語彙体系であれば、「語彙相互の関係性の明示」という、これまでにない特徴をもった語彙提示ができる。

また、さらに、「ある内容の文章を読ませて、それについてディスカッションをさせる」という課題を遂行する際には、オントロジーは、単なる語彙リストよりも、学習者の文生産に寄与できるはずである。

また、「所与のテキストから構築したオントロジー」は、「テキストの類型論」を考察する手がかりにもなりうる。例えば、一貫性をもつテキストの中で、オントロジーを構築することは、そのテキストの中では、トピックからサブトピックへ、どのように分岐していくかを示すことができる。これを利用し、「危うく死にかけた体験」「これまでで最も恥ずかしかったこと」などのテーマを与え、複数の言語でエッセイを書かせ、そこからオントロジーを構築すれば、言語間テキスト類型論を考察する手がかりにもなる。あるいは、特定の言語の母語話者とその言語を第二言語として学ぶ学習者に上記と同じ作業をさせれば、母語話者特有のテキスト構築のパターンを知ることもできよう。

5. 今後の課題

本アプリケーションは単にマークアップ作業を支援するだけではなく、すでにマークアップされたデータを入力することで統語構造やオントロジーの情報を手軽に図示することもできるため、言語学者の研究ツールとしての用途も考えられる。この立場からは、例えば任意の位置で枝を切って別の場所に移動する機能などのアイデアがあり、実装を検討している。

本アプリケーションは、手作業を簡略化する機能を搭載するのみで、自動処理の機能は搭載していない。例えば形態素解析等の自動化について、少なくとも日本語や英語といった主要な言語については搭載を検討している。また、オントロジーの付与についても、内部辞書を参照してサジェスチョンを出すような機能を実装できるか検討中である。

本アプリケーションは Flash で作成しているため、ウェブブラウザ上で動作させることができ便利な面もある一方で、データの保存に難点がある。この点は何らかの方法で解決する必要がある。

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, 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] 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.
[5] 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.
[6] 鈴木慎吾, 山崎直樹, 堀一成 “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”, 言語処理学会第 16 回年次大会発表論文集, (2010.3), pp.478-481.
[7] 野口正樹, 三好健太, 徳永健伸, 飯田龍, 小町守, 乾健太郎 “汎用アノテーションツール SLAT”, 言語処理学会第 14 回年次大会発表論文集, (2008.3), pp.269-272.
[8] 溝口理一郎 “オントロジー工学”, オーム社 (2005.1).
[9] Stanford Center for Biomedical Informatics Research, http://protege.stanford.edu/.