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

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 外国語教育研究機構　[†††]大阪大学 大学教育実践センター

1. はじめに

本稿では、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に現在開発中であるアプリケーションについて、その概略を述べる。

本稿の著者 3 名は、もともと旧大阪外国語大学で活動していた多言語同時処理プロジェクトのメンバーであり、現在は旧外大時代より蓄積してきた多言語平行コーパスを言語学および言語処理の分野に応用するための方法について研究を行っている。前稿 [1] では、その一環として開発を行った、コーパスに文構造情報のアノテーションを施すためのアプリケーションについて述べたが、今回紹介するアプリケーションはそれに続く試みである。

さて、オントロジーとは対象世界を概念化し、個々の概念間を関係づけることによって、暗黙の了解や前提知識を含んだ我々の現実世界に対する知識の構造を階層的に表現するものである。今回の試みは、これまでに蓄積してきた多言語コーパスにつき、そこに現れる語彙間を意味リンクで関連づけることにより、その背後にある知識構造をオントロジーとして表現しようというものである。

オントロジー構築ツールとしては法造 [2] や Protégé [3] 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ものであるが、今回のツール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。

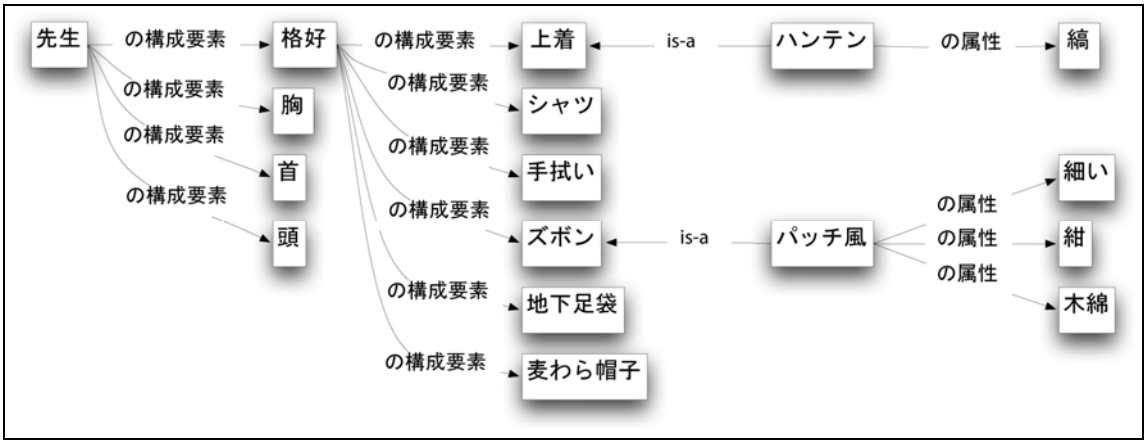
ところで、オントロジーとは本来、現実世界から抽象された「概念」を対象とするものであるが、本稿では「概念」ならぬ「語彙」、それもテキスト中に出現する「語彙」を対象としており、それはすでに具体的な存在である。したがって、さしあたり本稿が「オントロジー」として構築しようとしているものは、本来の意味における「オントロジー」そのものではなく、個々のオントロジーのインスタンスの関係を記述したもの、つまりオントロジーが現実世界において具現化したところの個別的事象間の関係をとらえたものである。これはひとえに本研究がコーパスを出発点としていることによるが、一方で言語形式と具体的事象との関係は言語学の主要なテーマの一つであり、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れが記述した世界のオントロジーを構築するのにも有益と思われる。

さて、前稿でも述べたが、我々が扱っているコーパスにはかなりマイナーな言語のものも含まれており、これらを対象としてオントロジーを構築するような場合、作業者の確保に苦労するという問題がある。したがって、ツールの作成にあたっては、該当言語の知識があるものであれば誰でも操作できるような、学習負担の少ないものとなるように特に留意している。また、外国語教育の立場から、言語教育実践に応用できるものも同時に目指している。

2. オントロジーの構築と埋め込み

例えば、次の文章（『窓ぎわのトットちゃん』の一節）をもとにオントロジーを構築するケースを考える。

> 校長先生は，こういって，一人の男の先生を，みんなに紹介した。トットちゃんは，つくづくとその先生を観察した。なにしろ，その先生の格好は，かわっていた。上着は縞のハンテンで，胸からは，メリヤスのシャツが，のぞいていて，ネクタイのかわりに，首には手拭いが，ぶら下がっていた。



【図 1】“先生”のオントロジー

そして，ズボンは，紺の木綿のパッチ風の細いのだし，靴じゃなくて，地下足袋だった。おまけに，頭には，少し破れた麦わら帽子をかぶっていた。

この文章に見える語彙（下線部）を拾い上げて、オントロジーを構築する方法によってそれらを関連づけてやると、図に示したような体系ができあがる（【図 1】）。

次に、できあがった体系のデータ化についてであるが、本ツールは構築された体系そのものをデータ化するのではなく、これらの情報を XML によって元のテキストに埋め込む方法をとる。そこで例えば上のケースであれば次のようにデータ化される [4]。

<su>…（略）…なにしろ，その<np id="n010">先生</np>の<np id="n011" partof="n010">格好</np>はかわっていた。<np id="n012" partof="n011">上着</np>は<np id="n013" attrof="n014">縞</np>の<np id="n014" is-a="n012">ハンテン</np>で，<np id="n015" partof="n010">胸</np>からは，<np id="n016" attrof="n017">メリヤス</np>の<np id="n017" partof="n011">シャツ</np>が，のぞいていて，ネクタイのかわりに，<np id="n018" partof="n010">首</np>には<np id="n019" partof="n011">手拭い</np>が，ぶら下がっていた。そして，<np id="n020" partof="n011">ズボン</np>は，<np id="n021" attrof="n023">紺</np>の<np id="n022" attrof="n023">木綿</np>の<np id="n023" is-a="n020">パッチ風</np>の<ajp id="n024" attrof="n023">細い</ajp>のだし，靴じゃなくて，<np id="n025" partof="n011">地下足袋</np>だった。おまけに，<np id="n026" partof="n010">頭</np>には，少し破れた<np id="n027" partof="n011">麦わら帽子</np>をかぶっていた。</su>

このようなデータ化は、コーパス中においてオントロジー情報を意味や統語情報といった他のメタ情報と併存させることができるという利点がある。

3. アプリケーションの詳細

3.1 概要

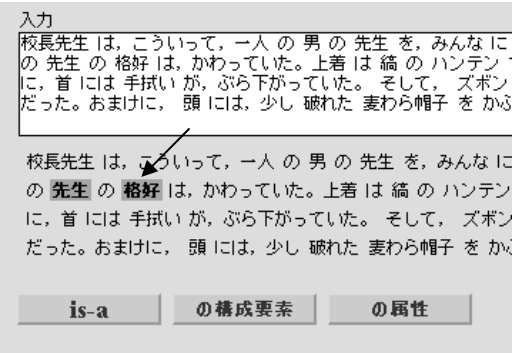
入力文からオントロジー情報を読み取り、その情報をもとのテキストに埋め込む作業を GUI ベースで行うことができる。動作環境は前回と同じく Adobe Flash であるので、Web システムとの親和性が高く、導入の敷居が低いのが特徴である。

なお、オントロジー構築の際には最初に形態素解析（単語分割）の必要があるが、本ツールは解析器を実装していないので、入力データはあらかじめ単語を空白で区切っておいたテキストを用いるものとする。実際には、形態素解析ソフトがある言語であればそれを用い、そうでない言語は前稿 [1] で発表したツールで大まかな句範疇をマークアップしたものを用いればよいだろう。

3.2　操作方法

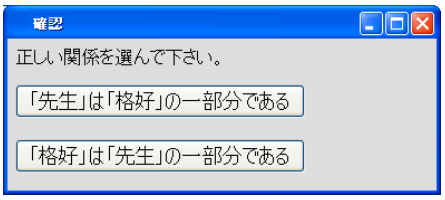
本ツールを使って入力文からオントロジーを構築する方法は以下の通りである。(全体画面を次頁【図 5】に示す。なお、画面は上から「入力エリア」、「作業エリア」、「出力エリア」の 3 つのエリアに分けられる。)

1. 入力エリアのテキストボックスに処理したい文を入力する。(→作業エリアに入力文が描画される)
2. 作業エリアで、関連づけの対象とする語彙を 2 つ選択する。それぞれの語彙はボタンになっていて、クリックにより選択することができる。(→選択された語彙がハイライト表示される【図 2】。)



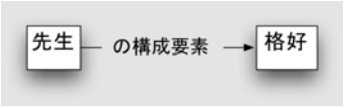
【図 2】語彙を選択

3. 関連づけの種類を「is-a」、「の構成要素」、「の属性」の 3 つから選び、ボタンを押す。ここでは「の構成要素」を選ぶ。(【図 2】)
4. 確認ダイアログがポップアップするので、関連づけの方向を選択する。(【図 3】)



【図 3】確認ダイアログ

5. 2 つの要素が関連リンクで結ばれた図が出力エリアに図示される。(【図 4】)
6. 2～5 を繰り返す。出力エリアには作成中のオントロジーが逐次表示され、それと共に対応する XML データも出力される。



【図 4】出力図

以上のように、ダイアログで確認しながら任意の 2 者の関係を指定していくだけなので、誰でも簡単に操作することができる。

3.3　補足説明

上記の他、付記すべき機能は以下の通り。

- 作業エリアの表示テキストで、すでにオントロジーのノードに用いられている語彙は太字で表示される。
- 出力文に複数のオントロジーが埋め込まれている場合はそれぞれのオントロジーが出力エリアに図示される。
- これまで説明した方法とは逆に、すでにオントロジー情報が埋め込まれたデータを出力エリアの XML データ出力部に入力することでオントロジーを描画することもできる。また XML データを直接編集することもできる。

4.　応用と課題

4.1　人文科学諸分野への応用

本ツールは人文科学の諸分野、特に言語学、文献学、歴史学、民俗学、人類学における資料作成ツールしての応用（下記参照）を考えている。

- 理論言語学：ある項目が、それが現れるテキスト内の意味ネットワークにおいてどのような位置を占めているかということとその項目の言語形式（定・不定、照応形式...など）との相関関係を調べる資料となる。
- 応用言語学
  - 言語教育：特定主題分野の「語彙マップ」の作成支援
  - 言語学習：文章読解支援ツール
- 文献学：
  - 白居易詩のオントロジー
  - 唐詩のオントロジー
  - 唐代知識人のオントロジー

4.2　今後の課題

今後の課題を挙げておく。

- GDA（およびそれに類する言語学的情報のためのマークアップ）との連携。
- OWL 形式の読み込みと書き出し。

謝辞

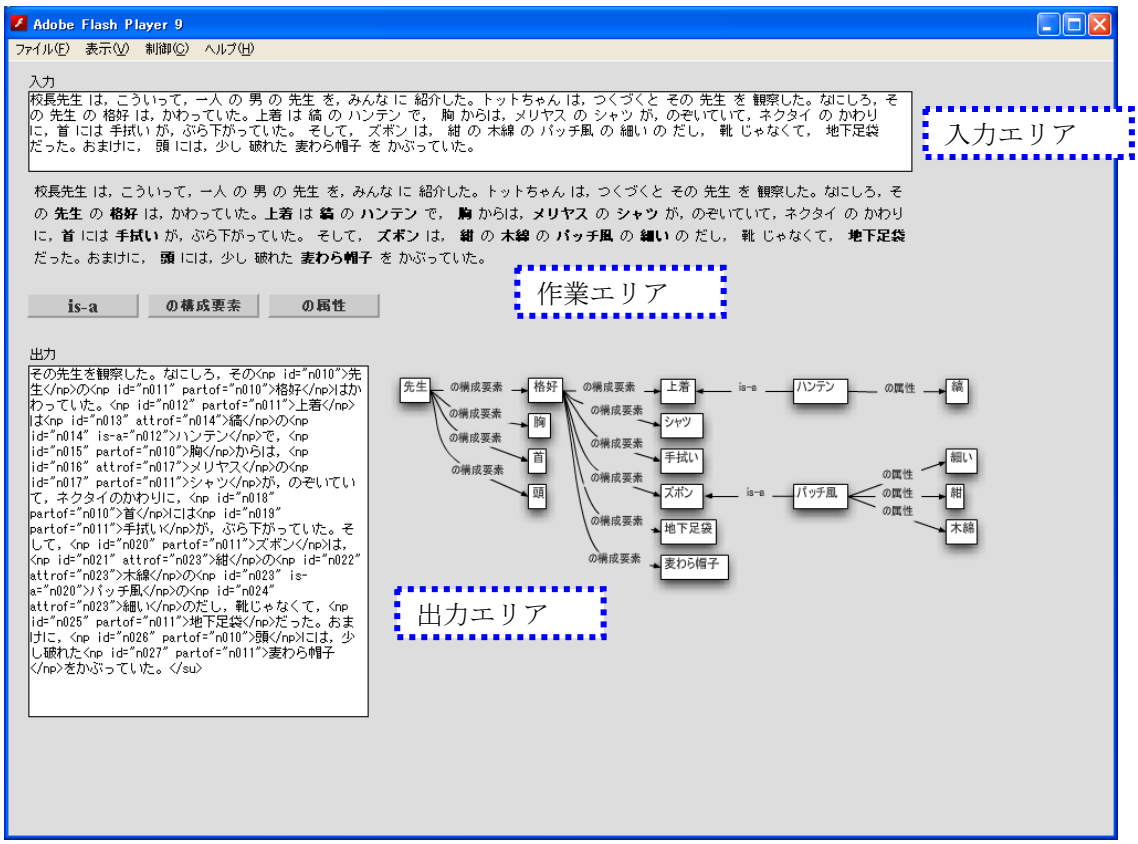
本研究は、科学研究費補助金　基盤研究（B）課題番号: 19300047『LCTL を含む多言語平行マルチメディア資源の構築と構造化方式の研究』（研究代表者: 堀　一成）の補助を受け推進したものである。

参考文献

溝口理一郎（著），人工知能学会（編集）『オントロジー工学』，オーム社, 2005.1.
溝口理一郎（編），古崎晃司, 來村徳信, 笹島宗彦, 溝口理一郎（著）『オントロジー構築入門』，オーム社, 2006.9.

注

[1] 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.
[2] 溝口理一郎研究室, http://www.hozo.jp/hozo/.
[3] Stanford Center for Biomedical Informatics Research, http://protege.stanford.edu/.
[4] ここでのデータ化は「大域文書修飾 Global Document Annotation (GDA)」を拡張した形式を採用している。GDA は http://i-content.org/gda/ を参照。なお本ツールは GDA のマークアップそのものは行わない。



【図 5】全体画面