エチオピア言語の Web サイト構築*

高橋　洋成
（筑波大学）
s025035@u.tsukuba.ac.jp

0　はじめに

我々のプロジェクトでは、エチオピアでのフィールド調査で得られた言語データ（語彙、文法、映像、地理情報など）をデータベース化する作業を進めている。現在、その作業の一環として、言語データを掲載した Web サイトの構築が進行中である。本稿は、構築中の Web サイトの内容および方向性を説明するものである。

1　Web サイトの必要性

我々のプロジェクトで行っているエチオピア言語のデータベース化は、大きく分けて 2 つの作業を行っている。1 つは GIS を用いた世界言語地図の作成、もう 1 つは言語データの XML 化と検索・照会システムの開発である。

まず、GIS を用いた言語地図に関しては、従来のシステムでは、どこに、何を、どのように表示するかを決定するには管理者権限が必要であった。しかし、杉井 (2009) において地図の属性情報と分布域の選択機能が付加されたことにより、必要となる GIS レイヤを利用者が選択し、さらには指定した値を持つものだけを地図上に表示することが可能になった。

一方、言語データの XML 化に関しては、高橋 (2008) が参考文献リストを BibTeXML 化し、現在は LaTeX で記述された言語データを Unicode に変換する作業が進められている。

ここで、2 つの作業の架け橋が必要となる。現在の GIS 言語地図は利用者が属性情報を選択できるが、どこまで選択肢を設ければ良いだろうか。現在までに収集された言語データは地理情報のみならず、語彙、文法、映像など多岐に

*本稿は 2007 年度～2009 年度科学研究費基盤研究 (B)「オモ・クシ系少数言語の調査研究及び地理情報システムを用いたデータベース構築」代表：乾秀行（山口大学）（研究課題番号：19401023）による研究成果の報告である。

渡り、しかも膨大な蓄積がある。それら全てを属性情報として一度に組み込むことは現実的ではない。

そこで、まず蓄積された言語データを Web サイトとして公開することにした。言語データを XML 化していくと同時に、Web ページに変換することは容易である。また、後で触れるように、Web ページに寄せられたフィードバックを XML データに反映させることも難しくない。そして、GIS で扱う属性情報も少しずつ調整していく。このように Web サイトをハブとすることで、互いの作業の成果、もしくは外部からのフィードバックに柔軟に対処し、データベースの精度を向上させることが期待できる。

2 Web サイトの構成

Web サイトの構築にあたっては 2 つのポイントがある。1 つは言語データの構成、もう 1 つは言語データの参照方法である。

2.1 言語データの構成

まず、言語データの構成に関して説明する。ここで述べる構成とは、言語データを掲載する際に、どのような項目を、どの程度まで記述するかの目安である。現在、プロジェクトメンバーである二ノ宮氏の提案に従い、1 つの言語に対し次のような構成で記述を行っている。

人口・系統・別称
: その言語の話者数、言語系統、呼び名などの情報である。これらは SIL (Summer Institute of Linguistics) が運営する Ethnologue[1] とリンクすることを考えている。Ethnologue では ISO 639-3: 2007 に基づく言語識別子が用いられており、コンピュータ処理しやすいという利点を持つ。だが一方で、言語と方言の境界をどこに定めるかといった分類上の問題があることも乾 (2008) で指摘されている。我々の言語データと Ethnologue のデータを照合していく作業は、言語の系統関係を整理することにもつながる。

音声・音韻
: 言語の音声学的特徴や音韻構造（特に音素目録）を提示し、具体的な例を挙げる。

語彙
: 調査済みの語彙を可能な限り提示する。また、特に親族名称と数詞をとり上げ、比較言語学研究のきっかけとなるような構成を目指す。

[1] http://www.ethnologue.com/web.asp

文法

名詞・動詞形態論の概要、語順および統語論的な特記事項を提示する。名詞形態論については性・数・格、動詞形態論については時制・法・相を想定している。もっとも、文法範疇の選定や分量に関しては対照言語学的な見地からの調整も必要となろう。

テキスト・会話

これらは言語コーパスとして利用できる。クシ・オモ系の少数言語のコーパスは、現在の言語の姿を保存するという意味でも非常に重要である。

参考文献

言語に関する主要な参考文献の一覧である。ハイパーリンクを用いることで快適な文献参照を行うことができる。

画像・動画

言語に関する画像や動画などを掲載する。音声を含む総体的な言語データとしても、また地理・文化を包括的に捉えるという点でも、重要なデータである。

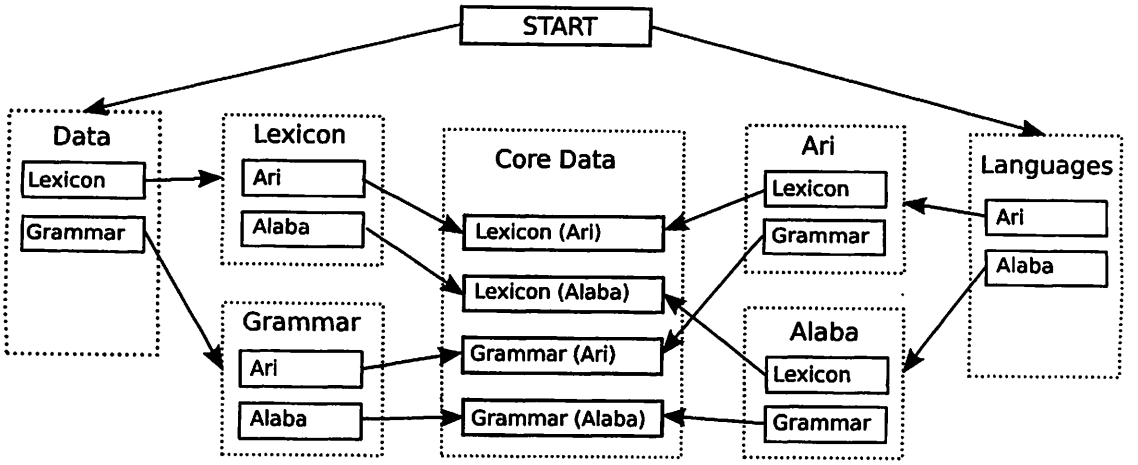
また、これらのデータ項目の数点を GIS と連携してマッピングを行う。

2.2 言語データの参照方法

本節で述べる参照方法とは、ナビゲーションと言い換えても良い。前節で言語データの構成項目を挙げたが、ある項目を参照するとき、目的によって探索の道筋を変更した方が良い場合がある。

例えば、アーリ語を体系的に調べたいという利用者がいたとする。そのような利用者に対するナビゲーションしては、まず「言語一覧」の中からアーリ語を選択させ、その中に音韻、語彙、文法、テキストなどの項目を配置するのが自然であろう。また、別の利用者は、ある語彙を持つ言語にはどのようなものがあるかを、言語横断的に調べたいと考えているかもしれない。そうした利用者に対しては、まず「項目一覧」の中から「語彙」を選択させ、その中にアーリ語、アラバ語などの言語名を配置するのが望ましい。

これらはナビゲーションページを設けることで解決できる。次の図は、Web サイトのトップページから言語データ（Core Data）に至るまでの道筋を表したものである。1 つの言語を体系的に調べたい利用者は右の「Languages」へ至る道を、あるデータ項目に関して言語を横断的に調べたい利用者は左の「Data」へ至る道を選べば良い。



なお、検索エンジンの導入という方法もありうる。そうすれば、フォームを用いた柔軟な検索も可能になるだろう。しかしながら、言語名の一致・不一致、Unicode 文字の入力手段、および Unicode の正規化に関する問題[2]など解決すべき問題が多いため、今回は導入を見送り、将来の課題としたい。

3 XML データベースとの関連

Web ページを作成する際、XML データベースとの連携も視野に入れておく。本節はその一例として RDFa の利用を検討する。

RDF (Resource Description Framework) とは、あるデータに関する情報（メタデータ）を記述するための枠組みである。RDF に従い、データとデータの関係を記述することで、特定のアプリケーションに依存しないデータベースを構築できる。RDFa とは、XHTML の属性（attribute）構文を利用して RDF 表現を行うというものである。

まず、次に挙げるのは通常の XHTML で記述された参考文献である。

```
<ol xmlns="http://www.w3.org/1999/xhtml">
  <li>
    Almagor, Uri,
    "Name-Oxen and Ox-Names among the Dassanetch of
       Southwest Ethiopia,"
    Paideuma 18 (1972) 79-96.
  </li>
  <li>
```

[2]例えば、「が」という文字に対し、合成済み文字「が」(U+304C) であるのか、それとも「か」(U+304B) と濁点 (U+3099) の結合文字列なのか、という問題である。処理プログラムは両者を同一の文字と扱わねばならない。

```
    Adams, Bruce,
    "A Tagmemic Analysis of the Wolaitta Language,"
    University of London,
    1984
  </li>
</ol>
```

次に、この XHTML に対し、RDFa を用いてメタデータを埋め込んだ例を挙げる。

```
<ol xmlns="http://www.w3.org/1999/xhtml"
    xmlns:bibtex="http://bibtexml.sf.net/"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  >
  <li typeof="bibtex:phdthesis">
    <span property="bibtex:author">Adams, Bruce</span>,
    "<span property="bibtex:title">A Tagmemic Analysis of
       the Wolaitta Language</span>,"
    <span property="bibtex:school">University of London</span>,
    <span property="bibtex:year" datatype="xsd:gYear">1984</span>
  </li>
  <li typeof="bibtex:article">
    <span property="bibtex:author">Almagor, Uri</span>,
    "<span property="bibtex:title">Name-Oxen and Ox-Names among
       the Dassanetch of Southwest Ethiopia</span>,"
    <span property="bibtex:journal">Paideuma</span>
    <span property="bibtex:volume">18</span>
    (<span property="bibtex:year" datatype="xsd:gYear">1972</span>)
    <span property="bibtex:pages">79-96</span>.
  </li>
</ol>
```

先に挙げたシンプルな XHTML と異なり、各項目がどのような書誌情報であるかを BibTeXML の要素タイプを用いて表現している。「そのデータが何を表しているのか」が明確になれば、高橋 (2008) が作成した文献データとリンクさせ、情報をアップデートしていくことが容易になる。

作業予定としては、まず先に挙げたようなシンプルな形で Web ページを記述していく。次に、必要に応じて RDFa などを用いてメタデータを埋め込んでい

く。メタデータによってデータと他のデータとの関係が分かれば、検索時に関連データを一緒に提示したり、データの修正を他の関連データにも反映させやすくなる。

4 おわりに

本稿は、エチオピア言語情報の Web サイト構築の内容、および方針を説明した。実際に運用段階に入れば、軌道修正が必要になる場面もあるだろう。そうした事態に備え、本稿は柔軟性に富んだ設計を目指している。

【参考文献】

Ben Adida et al (eds.) 2008 *RDFa in XHTML: Syntax and Processing.* W3C Recommendation. `http://www.w3.org/TR/2008/REC-rdfa-syntax-20081014/`

Davis, M. and K. Whistler 2008 *Unicode Character Database.* The Unicode Consortium. `http://www.unicode.org/Public/5.1.0/ucd/UCD.html`

Phillips A. and M. Davis (eds.) 2006 *RFC 4646: Tags for Identifying Languages.* `http://www.ietf.org/rfc/rfc4646.txt`

Phillips A. and M. Davis (eds.) 2006 *RFC 4647: Matching of Language Tags.* `http://www.ietf.org/rfc/rfc4647.txt`

乾秀行 2008「GIS を用いたデータベース構築に向けて」乾秀行（編）『オモ・クシ系少数言語の調査研究及び地理情報システムを用いたデータベース構築（Cushitic-Omotic Studies 2007）』1-8.

杉井学 2009「属性情報の地図分布を解析する GIS 構築」乾秀行（編）『オモ・クシ系少数言語の調査研究及び地理情報システムを用いたデータベース構築（Cushitic-Omotic Studies 2008）』1-4.

高橋洋成 2008「XML を利用した文献データベースの構築」乾秀行（編）『オモ・クシ系少数言語の調査研究及び地理情報システムを用いたデータベース構築（Cushitic-Omotic Studies 2007）』13-29.