フィールド言語学者のための電子辞書の開発*

山 田 久 就

1．はじめに

本稿の目的は筆者が開発を行っている電子辞書（仮名 MDict）について概略を示すことにあります。この電子辞書はフィールドワークをして集めた語彙を整理して、簡単に検索ができ、次の調査に効率よく利用できるようにすることを目指して開発を進めています。

筆者はアバール語の研究を行っていて、かなり前からアバール語の語彙を名詞、形容詞、動詞に分け XML 形式の辞書ファイルを作成しています[1]。そのため、この辞書ファイルに対して検索などの処理するプログラムが必要であったので、2000 年に最初の電子辞書を作成しました。当初の目的は、筆者自身が使用するためだけであったので、辞書ファイルを検索する機能だけのものでした。その後は多少の修正などを行っていましたが、2002 年に他者が使用するために提供する可能性を考慮して、辞書ファイルを作成して語彙を追加する機能を加えました。その後は開発を全く止めていましたが、2007 年度に入って開発を再開することを決め、時間を見つけて、多少の修正を行っています。なるべく早い時期に一般公開を行いたいと考えています。この電子辞書は Sun Microsystems が中心に開発を行っているプログラミング言語 Java を用いて開発を行っています。開発を始めた頃、筆者は Windows、MacOS、Linux などの Unix 系の OS といったいろいろな OS を利用していたので、いろいろな OS で同じプログラムを動かすことができるプログラミング言語 Java を採用しました。また、アバール語はキリル文字で表記するので、日本語で使用する文字とキリル文字を同時に表示するのに Java は適していることも Java を選んだ理由の一つです。2007 年 10 月 19 日現在、Java の最新のバージョンは Java2 Standard Edition 6 であり、Windows XP 上で javac 1.6.0_03 でコンパイルしたものを使っています。Java のプログラムを動かすには、Java の実行環境が必要です。Java の実行環境はいろいろな OS 用のものが無料で入手することができます。Windows 用の Java 実行環境は Sun Microsystems が開発していて、サイト［1］から入手することができます。

本稿の構成は次の通りです。2 節で電子辞書を動かすのに必要なファイルの構成について説明します。3 節では、電子辞書の使用方法について説明します。そして、4 節では XML 形式の辞書ファイルの構造について説明します。

2．ファイルの構成

電子辞書を実行するために必要なファイルについて説明します。電子辞書本体は mdict.jar という名前の実行ファイルです。mdict.jar の他に、設定に関するファイルが入っている resource という名前のフォルダと辞書ファイルが入っている dictionary という名前のフォルダが必要です。フォルダ resource は mdict.jar と同じフォルダに入れておく必要があります。フォルダ resource の中には preference.xml という名前のファイルと sort.txt という名前のファイルがあります。preference.xml はメニューなどで使用するフォント名やメニューなどで使われている項目の名前などが書かれています。このファイルは XML 形式のファイルであり、ファイルの内容を書き

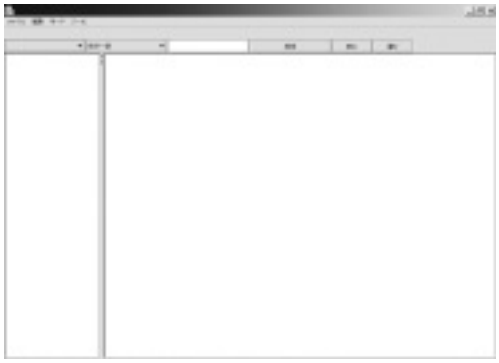
換えることによって使用するフォントを変えたり、メニューなどを別の言語で表示したりすることができます。sort.txt は単語をソートする順番を決定しているファイルです。ディフォルトではアバール語をソートする順番になっているので、アバール語以外の辞書を使用する場合はこのファイルを書き換える必要があります。フォルダ dictionary の中には品詞ごとの辞書ファイル noun.xml、verb.xml、adjective.xml と dtd というフォルダを置きます。noun.xml、adjective.xml、verb.xml はそれぞれ名詞、形容詞、動詞の辞書ファイルです。フォルダ dtd には nou.dtd、adjective.dtd、verb.dtd という名前のファイルがあり、ファイル nou.dtd、adjective.dtdverb.dtd はそれぞれ辞書ファイル noun.xml、adjective.xml、verb.xml の構造を定めているファイルです。

3．電子辞書の使用法

3.1　起動

電子辞書 MDict を起動するためには mdict.jar をダブルクリックします。MDict が起動すると図 1 のようなになります。

図 1



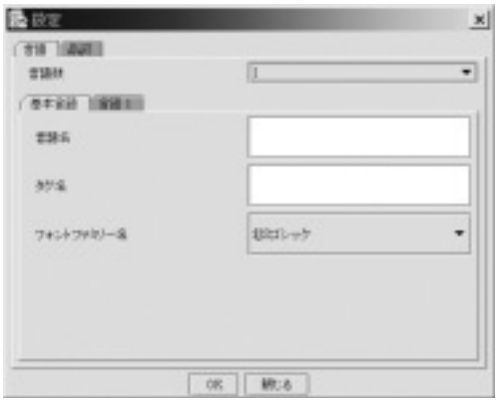
メインメニューは「ファイル」、「編集」、「モード」、「ツール」からなっています。メニュー「ファイル」の下にはサブメニュー「新しい辞書を作成」、「辞書ファイルをインポート」、「保存」、「終了」があります。メニュー「編集」の下にはサブメニュー「元に戻す」、「切り取り」、「コピー」、「張り付け」があります。メニュー「モード」の下にはサブメニュー「検索モード」、「登録モード」があります。メニュー「ツール」の下にはサブメニュー「設定」があります。ツールバーは二段になっています。上から第一ツールバー、第二ツールバーと呼ぶことにします。最初に起動した状態では、第一ツールバーには何も表示されてなく空の状態です。辞書ファイルがセットされた後は、第一ツールバーに品詞の名前が並びます。最初に起動した状態では、辞書ファイルがセットされていませんので、メニュー「ファイル」の下にあるサブメニュー「新しい辞書を作成」を使って空の新しい辞書ファイルを作成するか、メニュー「ファイル」の下にあるサブメニュー「辞書ファイルをインポート」を使って既存の辞書ファイルをインポートする必要かあります。

3.2　辞書ファイルのインポート

筆者が作成し使用しているアバール語の辞書を例に辞書ファイルのインポートについて説明し

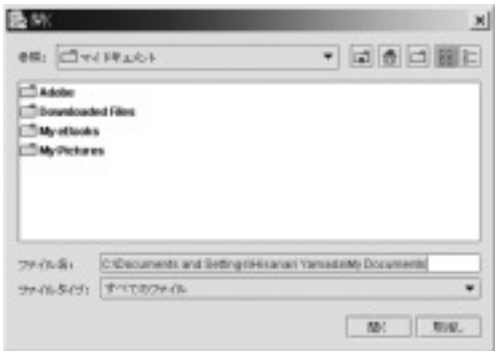
ます。辞書ファイルをインポートするには、メニュー「ファイル」の下にある「辞書ファイルをインポート」を選択します。そうすると、図 2 のようなダイアログが現れます。

図 2



このダイアログでは辞書ファイルの入っているフォルダを選択します。辞書ファイルが入っている dictionary というフォルダを探して選択します。次に、図 3 のダイアログが現れます。

図 3



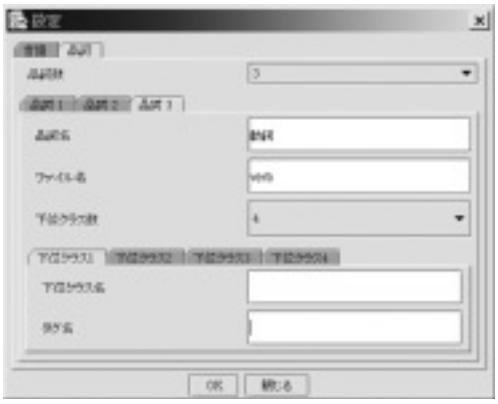
最初に、プルダウンメニュー「言語数」を選択します。「言語数」とは見出し語となる言語（基本言語と呼びます）に対して翻訳が付けられている言語の数を表しています。筆者のアバール語の辞書では、基本言語であるアバール語に対してロシア語と日本語で翻訳を付けているので、「言語数」は 2 となります。「言語数」を 2 に選択すると、「基本言語」、「言語 1 」という名前のタブの他に「言語 2 」という名前のタブが加わります。次に、「基本言語」、「言語 1 」「言語 2 」における「言語名」、「タグ名」、「フォントファミリー名」を設定します。「言語名」とは電子辞書の画面で使われているそれぞれの言語の名前です。筆者は日本語で名前を付けますが、他の言語で名前を付けることもできます。「タグ名」とは辞書ファイルで使用しているそれぞれの言語に対する名称です。筆者のアバール語の辞書ファイルではアバール語、ロシア語、日本語に対してそれぞれ

avar、rus、jp という名称を使用しています。「フォントファミリー名」は検索結果の画面でそれぞれの言語を表示するのに使用するフォントファミリーの名称です。これに基づいて、タブ「基本言語」のテキストボックス「言語名」にアバール語、テキストボックス「タグ名」に avar と書き込み、プルダウンメニュー「フォントファミリー名」をたとえば Times New Roman にします。次に、タブ「言語 1 」をクリックして、テキストボックス「言語名」にロシア語、テキストボックス「タグ名」に rus と書き込み、プルダウンメニュー「フォントファミリー名」をたとえば Times New Roman にします。次に、タブ「言語 2 」をクリックして、テキストボックス「言語名」に日本語、テキストボックス「タグ名」に jp と書き込み、プルダウンメニュー「フォントファミリー名」をたとえば MS P 明朝にします。これで言語に関する設定は終わりです。次に「品詞」タブをクリックします。図 4 のようになります。

図 4

最初に、プルダウンメニュー「品詞数」を選択します。筆者のアバール語の辞書は名詞、形容詞、動詞からなっているので、「品詞数」3 に選択します。そうすると、タブ「品詞 1 」に「品詞 2 」、「品詞 3 」という名前のタブが加わります。それぞれの品詞に対して、「品詞名」、「ファイル名」、「下位クラス数」を設定します。「品詞名」とは電子辞書の画面に表示される品詞名です。何語で設定してもかまいません。「ファイル名」とは、辞書ファイルの名前から拡張子の部分（すなわち、.xml）を省いたものです。「下位クラス数」とは辞書ファイルでそれぞれの品詞の単語を何個の下位クラスに分類しているかを意味します。筆者のアバール語の辞書ファイルでは、名詞と形容詞には下位クラスを設定していませんが、動詞は自動詞、他動詞、与格動詞、位格動詞の四つに分類しています。ます、「品詞 1 」のテキストボックス「品詞名」に名詞、テキストボックス「ファイル名」に noun と書き込みます。次に、タグ「品詞 2 」をクリックして、テキストボックス「品詞名」に形容詞、テキストボックス「ファイル名」に adjective と書き込みます。次に、タグ「品詞 3 」をクリックします。テキストボックス「品詞名」に動詞、テキストボックス「ファイル名」に verb と書き込みます。それから、プルダウンメニュー「下位クラス数」を 4 にします。すると、図 5 のようになり、「下位クラス 1 」、「下位クラス 2 」、「下位クラス 3 」、「下位クラス 4 」という名前のタブが現れます。

図 5



それぞれのタブで「下位クラス名」と「タグ名」を設定します。タブ「下位クラス 1 」、タブ「下位クラス 2 」、タブ「下位クラス 3 」、タブ「下位クラス 4 」のテキストボックス「下位クラス名」、テキストボックス「タグ名」をそれぞれ自動詞と intr、他動詞と tr、与格動詞と dat、位格動詞と loc にします。「タグ名」とは辞書ファイルで使用しているそれぞれの下位クラスに対する名称です。最後に、ボタン「OK」をクリックします。これで、辞書ファイルのインポートが行われます。

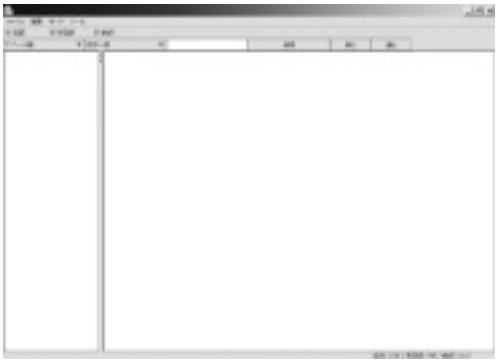
3.3　新しい辞書の作成

空の新しい辞書ファイルを作るには、メニュー「ファイル」の下にあるサブメニュー「新しい辞書を作成」を選択します。そうすると、図 3 のダイアログが現れます。後は、自分が望む設定を行います。

3.4　検索モード

検索は検索モードで行います。検索モードの画面は図 6 のようになります。

図 6



検索モードの第一ツールバーには品詞の名前が並んでいます。第二ツールバーには左から二つプ

ルダウンメニュー、テキストボックス、「検索」ボタン、「戻る」ボタン、「進む」ボタンがあります。左側のプルダウンメニューには言語名が並んでいます。右側のプルダウンメニューは「前方一致」、「後方一致」、「完全一致」、「含む」、「正規表現を使う」からなっています。

検索を行うには、第一ツールバーから検索したい品詞を選択し、第二ツールバーの左側のプルダウンメニューから検索したい言語を選択し、第二ツールバーの右側のツールバーから検索方法を選び、テキストボックスに単語あるいは単語の部分を書き込み、「検索」ボタンをクリックします。第一ツールバーの検索したい品詞は複数を選択することができます。検索の結果は図7のようになります。

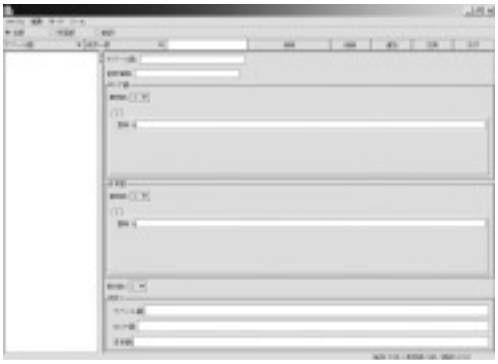
図 7

第二ツールバーにある「戻る」ボタンをクリックすると一つ前に検索した単語が現れて、「進む」ボタンをクッリクすると一つ後に検索した単語が現れます。

3.5　登録モード

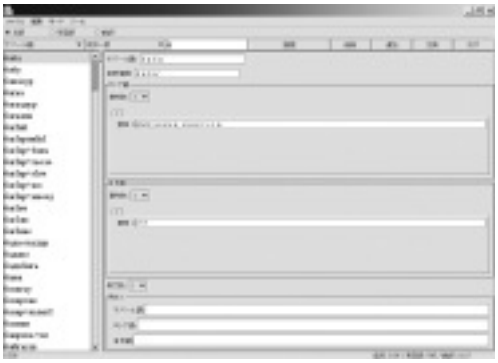
単語を追加したり、削除したり、内容を変更したりするのは登録モードで行います。検索モードから登録モードへ移るには、メニュー「モード」の下にはサブメニュー「登録モード」を選択します。登録モードの画面は図8のようになります。

図 8



登録モードから検索モードに戻るにはメニュー「モード」の下にはサブメニュー「検索モード」を選択します。登録モードの第一ツールバーには品詞の名前が並んでいます。検索モードでは検索するために複数の品詞を選択することができるのに対して、登録モードではどれか一つの品詞しか選択できないようになっています。登録モードの第二ツールバーは左から二つのプルダウンメニュー、テキストボックス、「検索」ボタン、「削除」ボタン、「追加」ボタン、「交換」ボタン、「クリア」ボタンがあります。二つのプルダウンメニューは検索モードと同じです。「検索」ボタンは単語を検索するためにあります。検索の結果は図 9 のようになります。

図 9



「クリア」ボタンは図 9 のような画面を図 8 のような初期画面に戻します。単語を追加するには「追加」ボタンを用います。初期画面から必要な項目を埋めていって単語を追加することもできるし、図 9 のような検索結果からいくつかの項目を書き直して単語を追加することもできます。単語を追加するためには最低限見出し語が埋められている必要があります。単語を削除するには図 9 のように検索結果で単語が選択されている状態で「削除」ボタンをクリックします。単語を削除すると復活させることはできません。単語の登録内容を変更するには「交換」ボタンを用います。「検索」ボタンで単語を検索して選択した状態で必要な項目を書き直して「交換」ボタンをクリックします。登録内容を変更すると元に戻すことはできません。辞書の項目を追加したり削除したりして変更した際はメニュー「ファイル」の下にあるサブメニュー「保存」を選択して変更内容を保存する必要があります。

4．辞書ファイルの構造

辞書ファイルは XML 文書であり、図 10、図 11 のような構造をしています。

図 10

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE verb SYSTEM "dtd/verb.dtd">
<verb>
<item>
<avar>word 1</avar>
<phonetic>pronounciation 1</phonetic>
<intr>
<rus><me>meaning 1</me></rus>
<jp><me> meaning 2</me></jp>
<ex>
<avar>example 1</avar>
<rus><me> meaning 3</me></rus>
<jp><me> meaning 4</me></jp>
</ex>
</intr>
<tr>
<rus><me>meaning 5</me></rus>
<jp><me> meaning 6</me></jp>
<ex>
<avar>example 2</avar>
<rus><me> meaning 7</me></rus>
<jp><me> meaning 8</me></jp>
</ex>
</tr>
</item>
</verb>
```

メニュー「ファイル」のサブメニュー「新しい辞書を作成」を使って新しい空の辞書を作成して、単語を追加していく場合には辞書ファイルを自分で作成する必要はありませんが、メニュー「ファイル」のサブメニュー「辞書ファイルをインポート」を使って辞書ファイルをインポートするためには語彙を整理しているファイルを図 10、図 11 のような辞書ファイルに作り替える必要があります。図 10、図 11 はそれぞれ筆者のアバール語の辞書の名詞に関するファイル noun.xml と動詞に関するファイル verb.xml の構造を示しています。XML 文書とは簡単に言うと開始タグ＜xxx＞と終了タグ＜/xxx＞で囲まれた範囲にテキストや別の開始タグと終了タグのペアーが入っている文書です。xxx には任意の文字列を入れることができます。XML についての公式の情報はサイト［2］で得ることができます。また、XML についてより詳しく知りたい場合は文献［1］などを参照してください。図10の 1 行目と 2 行目は一種の前書きです。encoding＝"yyy" の yyy にその文書で使っている文字コードが書かれます。筆者のアバール語の辞書では utf-8 です。XML 文書の実体（XML インスタンスと呼ばれる）は 3 行目の＜noun＞から始まって最終行の＜/noun＞で終わります。一つの単語に関する情報は＜item＞と＜/item＞の間にあります。図 10 には＜item＞と＜/item＞のペアーが二つあるのでこの辞書ファイルには二つの単語についての情報が含まれています。図 10 の斜体字の部分が実際の情報です。たとえば、＜avar＞*word 1*＜/

図 11

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE verb SYSTEM "dtd/verb.dtd">
<verb>
<item>
<avar>word 1</avar>
<phonetic>pronounciation 1</phonetic>
<intr>
<rus><me>meaning 1</me></rus>
<jp><me> meaning 2</me></jp>
<ex>
<avar>example 1</avar>
<rus><me> meaning 3</me></rus>
<jp><me> meaning 4</me></jp>
</ex>
</intr>
<tr>
<rus><me>meaning 5</me></rus>
<jp><me> meaning 6</me></jp>
<ex>
<avar>example 2</avar>
<rus><me> meaning 7</me></rus>
<jp><me> meaning 8</me></jp>
</ex>
</tr>
</item>
</verb>
```

avar＞の部分はこの単語の見出し語（アバール語での表記）が *word 1* であることを示しています。＜phonetic＞*pronounciation 1*＜/phonetic＞は音声情報（発音情報）が *pronounciation 1* であることを示しています。＜rus＞と＜/rus＞の間には＜me＞*meaning 1*＜/me＞と＜me＞*meaning 2*＜/me＞が含まれていますが、これはロシア語に訳す場合二つの意味に分けることができ、第一の意味が *meaning 1* であり、第二の意味が *meaning 2* であることを示しています。同様に、＜jp＞と＜/jp＞の間には＜me＞*meaning 3*＜/me＞と＜me＞*meaning 4*＜/me＞が含まれていますが、これは日本語に訳す場合二つの意味に分けることができ、第一の意味が *meaning 3* であり、第二の意味が *meaning 4* であることを示しています。＜ex＞と＜/ex＞の間に例文（*example 2* や *example 3*）とそのロシア語訳（*meaning 9* や *meaning 11*）、日本語訳（*meaning 10* や *meaning 12*）が書かれています。＜ex＞と＜/ex＞のペアーが二つあるので二つの例文を含んでいます。

図 11 の構造は図 10 の構造と少し違っています。その構造の違いは筆者のアバール語の辞書において名詞には下位クラスを設定していないのに対して、動詞では四つの下位クラス（自動詞、他動詞、与格動詞、位格動詞）を設定していることによります。図 11 には＜item＞と＜/item＞のペアーが一つしかないので一つの単語に関する情報だけが含まれています。＜item＞と＜/item＞は＜intr＞と＜/intr＞のペアーと＜tr＞と＜/tr＞のペアーを含んでいます。これはこの動詞には自動詞としての用法と他動詞としての用法があることを示しています。＜intr＞と＜/intr＞の間、＜tr＞と＜/tr＞の間には図 10 で説明したような情報が含まれています。

以上が辞書ファイルの説明ですが、辞書ファイルにどのようなタグが使われているのか、タグと実際の名前の対応（たとえば、タグの名前「rus」と実際の名前「ロシア語」、タグの名前「intr」と実際の名前「自動詞」）は辞書ファイルをインポートするときに図 3 、図 4 、図 5 で示されてい

る設定ダイアログで指定します。

5．ソートファイル

ソートファイルはアルファベットの順番を示したファイルです。たとえば、ある言語のアルファベットがｓ、ｂ、ｚ、ｒ、ｔ（大文字はＳ、Ｂ、Ｚ、Ｒ、Ｔ）の五つからなっていてこの順序でソートし、かつ小文字を大文字の前に持っていきたい時には、ソートファイルに次のように書いてください。

```
< s, S < b, B < z, Z < r, R < t, T
```

また、大文字を小文字の前に持っていきたい時は、ソートファイルを次のように書いてください。

```
< S, s < B, b < Z, z < R, r < T, t
```

最初の‘<’は必ず書いてください。ソートファイルで使える文字コードはUTF-8だけです。

6．おわりに

以上が、現在開発中の電子辞典の概要です。まだ、基本的な機能しかありませんが今後機能を追加していく予定です。また、実験的な段階ですので、いろいろな面でかなりの変更を行うことになります。しかし、なるべく早い時期に一般公開をできる状態に持って行きたいと考えています。

注

*　本稿は文部科学省科学研究費補助金(若手研究(B)、研究課題：『現地調査とデータベース作成によるアバール語の現状と変容に関する社会言語学的研究』、課題番号：17720076、研究代表者：山田久就、研究期間：2005-7年度）から助成を受けている研究の成果の一部である。

1．アバール語は北東コーカサス諸語（あるいはダゲスタン諸語）に属し、主にロシア連邦ダゲスタン共和国およびアゼルバイジャン共和国で話されています。

参 照 文 献

[1] 中山幹敏，奥井康弘（編著）『改訂版標準 XML 完全解説（上）』，技術評論社，2001 年

参照サイト（2007 年 10 月 19 日現在）

[1] http://www.java.com/ja/
[2] http://www.w3.org/TR/REC-xml