TR-IT-0313

TDMT 辞書ツール説明書 (英日版・ユーザ向け)

TDMT Dictionary Tool Users Manual (English to Japanese)

鷹尾 和享　　柏岡 秀紀　　白井 諭
Kazutaka TAKAO　Hideki KASHIOKA　Satoshi SHIRAI

1999 年 9 月 30 日

概要

TDMT の英日の翻訳に使われる辞書を効率的に管理するツールの利用方法について述べる。このツールは Windows95 で動作する。主な機能としては、単語登録、辞書検索、行単位の編集、角川類語新辞典の検索などがある。また、意味コードを自動推定する機能があるので、辞書管理者の負担を軽減することができる。

エイ・ティ・アール音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratories

目次

第１章

はじめに

1.1　システム概要

TDMT 辞書ツールは TDMT の英日の翻訳に使われる辞書を効率的に管理するツールである。TDMT 辞書ツールのシステム構成は図１の通りである。UNIX 上にある TDMT 辞書を Windows に転送する。１つの巨大な辞書ファイルに対して読み書きを行うと効率が悪いので、辞書ファイルを細かく分割して分割辞書として Windows のディスク上に配置し、それに対して単語の追加・修正等を行う。その際、必要に応じて角川類語辞典の CD を参照する。分割辞書はきりのいい時点で再結合し、UNIX に転送して TDMT で使用することとなる。

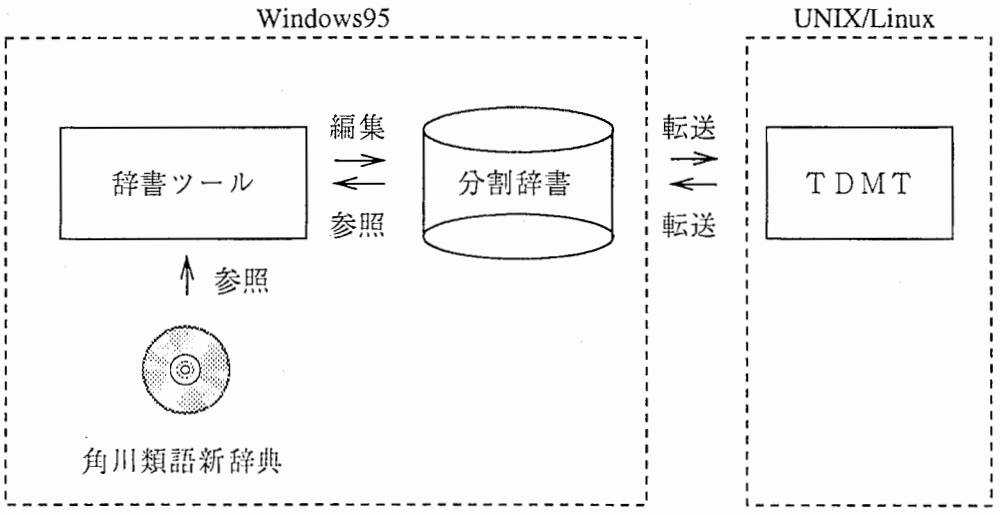

図１　システム構成

1.2 作業の流れ

ユーザの作業手順は以下のようになる。

(1) 辞書ツールのインストール。
(2) UNIX 上の TDMT 辞書を ftp 等で Windows に転送し、辞書分割ツールで分割する。
(3) 辞書ツールを使って、分割した辞書に対して単語の追加・修正等を行う。
(4) 分割辞書を再結合し、ftp 等で UNIX に転送する。

上記の(1)(2)はインストール時に1回だけ行えばよく、日常の作業では(3)(4)を繰り返して辞書ファイルの整備を行うことになる。

1.3 機能・特徴

TDMT 辞書ツールには以下のような機能・特徴がある。

- 単語を新しく登録したい場合、英語・日本語・品詞・意味コード・活用形を入力すれば、英日対訳辞書・英語意味コード辞書・英語形態素辞書にまとめて単語を登録できる。つまり、いちいち個別に登録する必要がない。
- 英語と日本語訳を入力すれば、品詞・意味コード・活用形を自動的に推定することができる。
- 未知語リストのファイルから次々に単語を読み込み、英語・日本語の欄に表示することができる。
- 既存の辞書から単語を検索することができる。
- 検索結果に対して辞書ファイルの1エントリ単位の編集をすることができる。（1エントリ＝1行）
- 検索結果から辞書の行を選んで辞書ファイルから削除することができる。
- 辞書ファイルに対して1エントリ単位で新しく追加することができる。
- 角川類語新辞典の検索をすることができる。したがって、類語辞典の検索結果を画面上に表示しながら意味コードの付与作業をすることができる。
- 辞書ファイルを細かく分割することによってファイル1個あたりのサイズを小さくし、

効率的に入出力を行うとともに、エディターの編集サイズの制限を回避する。

また、本ツールが対象とするTDMT辞書は以下の4つである。

english-to-japanese.lisp（英日対訳辞書）

ema-atr-sem-code.text（英語意味コード辞書）

ema-eng-morph-manual.dic（英語形態素辞書、手で編集用）

ema-eng-morph.dic（英語形態素辞書、タグから機械的に作成）

なお、ema-eng-morph.dicは参照専用であり、これに対する編集操作はできない。

第２章

インストール

2.1　必要環境

OS

日本語 Windows95

（Windows98、NT 4.0 での動作確認は行っていない）

CPU

Pentium 200MHz 以上を推奨

メモリ

Windows の必要分のほかに 3MB 程度必要

（搭載メモリで言うと 32MB 必要、できれば 64MB が望ましい）

ディスク容量

辞書ツールプログラム　1MB

辞書　130MB（35 万語の場合）

（高速なディスクが望ましい）

その他

角川類語新辞典 CD-ROM 版が必要。

UNIX 上の TDMT 辞書をやりとりするので、ネットワーク環境が必要。

Visual C++ 4.0 のランタイム環境が必要。

注：必要なのはランタイム環境のみである（再頒布可能）。具体的には

MFC40.DLL　MFC40LOC.DLL　MSVCRT40.DLL

が必要である。残念ながら、1999 年 9 月現在、インストールプログラムは用意していない。ただし、マシンによっては既に入っている可能性はある。

2.2 辞書ツールのインストール

(1) インストール先のディレクトリを新しく作り、そこへ全てのファイルをコピーしてください。なお、コマンドラインからディレクトリ名を入力する場合があるので、"Program files" 等ではなく、入力しやすい名前にすることを強くお勧めします。

例：mkdir C:¥DICTOOL

(2) 分割辞書用のディレクトリを新しく作り、さらにその下に4つの辞書用のディレクトリをそれぞれ作ってください。

例：mkdir C:¥DIV

mkdir C:¥DIV¥EJ … english-to-japanese.lisp 用

mkdir C:¥DIV¥ES … ema-atr-sem-code.text 用

mkdir C:¥DIV¥EM … ema-eng-morph-manual.dic 用

mkdir C:¥DIV¥EMA … ema-eng-morph.dic 用

(3) メモ帳等で DICTOOL1.INI と MICHIGO0.INI を編集してください。
両者同一内容です。2つとも編集してください。

[CD] … 角川類語新辞典関連の設定

zkana=Q:¥TITLE¥ZKANA … Q:を CD-ROM のドライブに直す

honmon=Q:¥RUIGO.DAT¥HONMON … 〃

cache=C:¥WINDOWS¥TEMP¥MIDASHI.TMP … ｷｬｯｼｭﾌｧｲﾙ1の場所と名前

cache2=C:¥WINDOWS¥TEMP¥MIDASHI2.TMP … ｷｬｯｼｭﾌｧｲﾙ2の場所と名前

[divdicdir] … (2)で作った分割辞書のディレクトリ（末尾に¥が必要）

ej=C:¥DIV¥EJ¥ … english-to-japanese.lisp 用

es=C:¥DIV¥ES¥ … ema-atr-sem-code.text 用

em=C:¥DIV¥EM¥ … ema-eng-morph-manual.dic 用

emRO=C:¥DIV¥EMA¥ … ema-eng-morph.dic 用

※キャッシュファイルとは、高速化のために CD-ROM の必要部分をディスク上に記憶しておくためのファイルで、辞書ツールが自動的に作成する。

(4) DICTOOL1.INI と MICHIGO0.INI を Windows ディレクトリに移動させてください。

※ Windows ディレクトリ：Windows95 を C:¥WINDOWS にインストールしていれば C:¥WINDOWS。わからない場合は MS-DOS プロンプトを開いて set と入力し、windir=の欄を見ればよい。

(5) 適宜 DICTOOL1.EXE や MICHIGO0.EXE のショートカットをスタートメニューやデスクトップに作るとよいでしょう。

DICTOOL1.EXE … 辞書ツール

MICHIGO0.EXE … 応用プログラム：未知語を学習する仕掛け

2.3 TDMT 辞書のセットアップ

(1) 辞書ファイルを Shift-JIS に変換（UNIX/Linux で）

例：

```
coco ¥*internal¥* ¥*sjis¥* < /xxx/english-to-japanese.lisp > /tmp/english-to-japanese.lisp
coco ¥*internal¥* ¥*sjis¥* < /yyy/ema-atr-sem-code.text > /tmp/ema-atr-sem-code.text
coco ¥*internal¥* ¥*sjis¥* < /zzz/ema-eng-morph-manual.dic > /tmp/ema-eng-morph-manual.dic
coco ¥*internal¥* ¥*sjis¥* < /zzz/ema-eng-morph.dic > /tmp/ema-eng-morph.dic
```

(2) ftp（ASCII モード）で Windows に転送

Windows で DOS 窓を開いて ftp を起動し、上記の 4 つのファイルを get する。

Windows 側の場所は次のステップの便宜のため C:¥DICTOOL がよい。(SPLDIC.EXE と同じ場所に入れておくと、ファイル名をいちいちフルパス名で入力する必要がない)

(3) 辞書分割ツール SPLDIC.EXE で辞書を分割（Windows で DOS 窓を開いて）

それぞれの辞書が 729 個のファイルに分割されます。

例：

```
cd C:¥DICTOOL
spldic ej        english-to-japanese.lisp        C:¥DIV¥EJ¥
```

```
spldic esem      ema-atr-sem-code.text        C:¥DIV¥ES¥
spldic emorph    ema-eng-morph-manual.dic     C:¥DIV¥EM¥
spldic emorph    ema-eng-morph.dic            C:¥DIV¥EMA¥
```

※末尾の ¥ を忘れないように。

※エラーログは#ERR というファイルに出力するので、エラーがあった場合は辞書ファイル（english-to-japanese.lisp 等）を編集し、分割ディレクトリのファイル（C:¥DIV¥EJ¥*.*等）を削除してから spldic をやり直すこと。

※作業が終われば english-to-japanese.lisp 等の 4 つの辞書ファイルは削除してよい。

第 3 章

操作説明

3.1　起動

角川類語新辞典の CD をセットし、C:¥DICTOOL の DICTOOL1.EXE を起動する。初回起動時には角川類語新辞典の CD からキャッシュファイルの作成を行うので、立ち上がるまで多少時間がかかる。2 回目からはキャッシュファイルから読み込むので、CD は不要で、立ち上がる時間も短い。なお、キャッシュファイルには著作権保護のためスクランブルをかけてあるので、そのままエディター等で読むことはできない。

3.2　初期メニュー

図 2　初期メニュー

[辞書登録]

単語登録のウインドウを開く

[検索]

辞書検索のダイアログボックスを開く

[検索結果]

検索結果のウインドウを開く

[RUI]

角川類語新辞典の検索のウインドウを開く

[終了]

辞書ツールを終了する

3.3 辞書登録

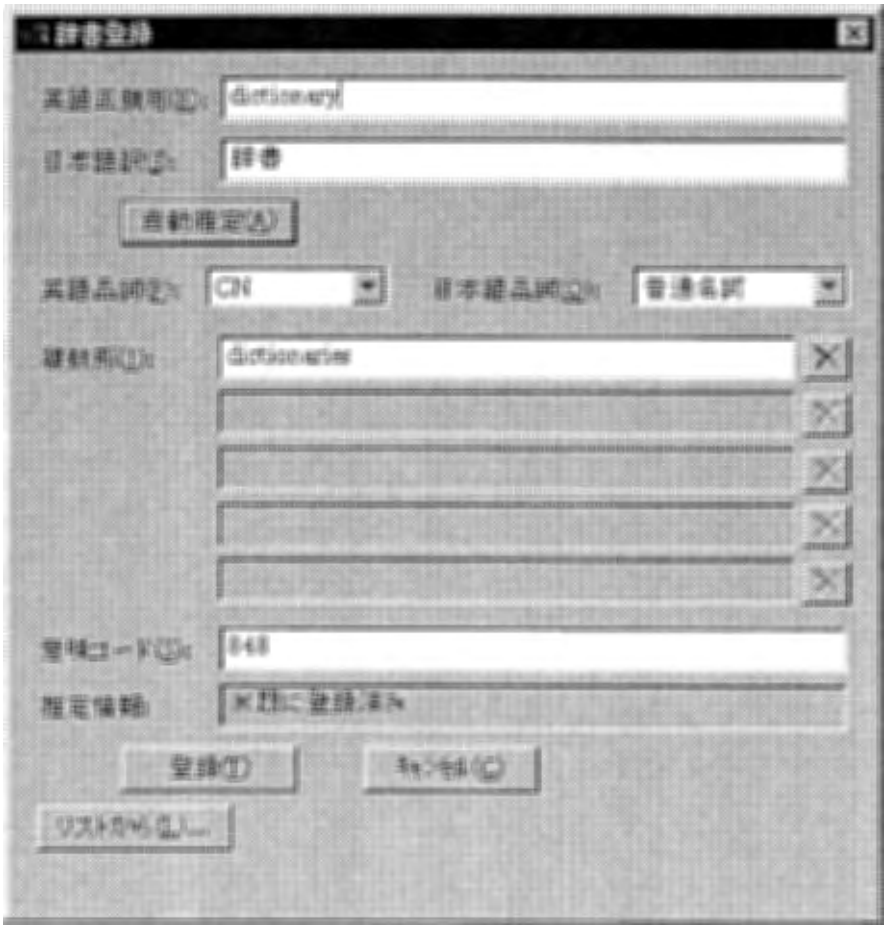

図3　辞書登録

辞書に新しく単語を登録します。

[英語正規形]

英語の単語を正規形で入力してください。

[日本語訳]

日本語の訳を入力してください。

[自動推定]

英語正規形と日本語訳を入力した後、このボタンを押すと、英日の品詞・活用形・意味コードを自動的に推定します。

[英語品詞]

英語の品詞を選択してください。

[日本語品詞]

日本語の品詞を選択してください。

[活用形（複数形など）]

英語の活用形を入力してください。項目は[英語品詞]の選択によって変わります。

その活用形が不要の場合は空欄にしてください。

[×]

活用形が不要の場合、これを押すことにより欄をクリアできます。

[意味コード]

意味コードを半角で入力してください。複数入力する場合は半角スペースで区切ってください。なお、""で囲む必要はありません。

[推定情報]

[自動推定]で推定した意味コードの根拠に関する情報が表示されます。

これが間違っている場合は、自動推定した意味コードも間違っていると判断できます。

[登録]

単語を辞書に登録します。

[キャンセル]

登録を中止します。

[リストから]

未知語リストのファイルから単語を次々と読み込んで表示する機能です。[自動推定]がなされた状態で表示されます。この場合、[登録]で辞書登録をして次の語に移り、[キャンセル]で登録せずに次の語に移ります。ファイルの形式は

英語正規形　タブ　日本語

を１行ずつ記述します。なお、英語正規形のみを記述した場合は、読み込んだ後に日本語を入力することになります。

【特殊な品詞の特別処理】

いずれも形態素辞書のみに出現するが、「英語正規形」の欄に入力してください。

● SYMBOL

・日本語不要

・意味コード不要

● 00

・日本語不要

・意味コード不要（既に登録済み）

・品詞を 00 とし、出現形を英語正規形の欄に入力

●^arabic

・日本語不要

・意味コード不要（既に登録済み）

・品詞を ^arabic とし、出現形を英語正規形の欄に入力

●^arabsym

・日本語不要

・意味コード不要（既に登録済み）

・品詞を ^arabsym とし、出現形を英語正規形の欄に入力

【推定情報の欄の表示内容の例】

●例１：日本語訳に「旅行会話」を入力し、「会話　※角川を参照」となった場合

「旅行会話」が辞書になかったので、「会話」で角川を参照し、その意味コードを表示

●例２：英語正規形に「travel　conversation」を入力し、「conversation　※主要な語を探し、自分自身を参照」となった場合

「travel conversation」が辞書になかったので、その中の主要な語を探して「conversation」で既存の意味コード辞書を参照し、その意味コードを表示

●例３：日本語訳に「もろさ」を入力し、「もろい　※角川を参照」となった場合

「もろさ」が辞書になかったので、語形を変化させて「もろい」で角川を参照し、その意味コードを表示

3.4　辞書検索

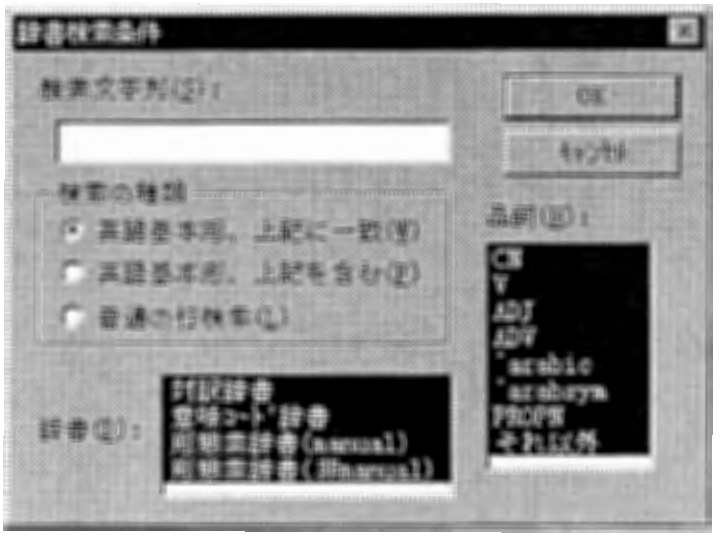

図４　辞書検索条件

分割辞書から語や文字列を検索します。

[検索文字列]

検索する語・文字列を入力してください。

[検索の種類]

・英語基本形、上記に一致

英語の基本形を指定し、検索を行います。基本形を正確に入力する必要がありますが、

非常に高速です。

・英語基本形、上記を含む

英語の基本形の一部を指定し、検索を行います。正確な入力でなくてもかまいませんが、低速です。

・普通の行検索

単純に行検索を行います。したがって、日本語やコメント部分も検索対象になります。低速です。

[品詞]

検索対象の品詞を少なくとも１つ選んでください。

[辞書]

検索対象の辞書を少なくとも１つ選んでください。

3.5　検索結果

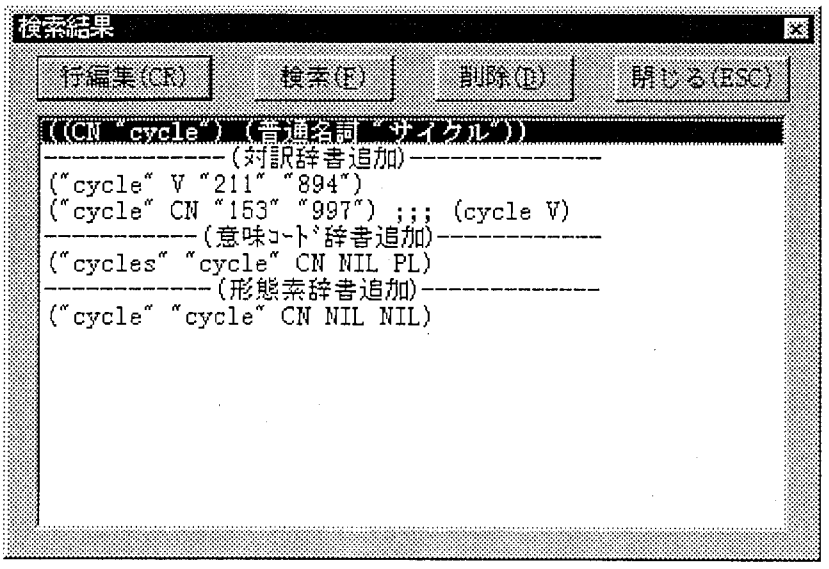

図５　検索結果

検索結果が表示されます。また、行単位の編集や、削除、新規追加もできます。

[行編集]

一覧の ListBox のうちの選択した行を編集することができます。

辞書行編集のダイアログボックスが表示されます。

[検索]

辞書検索条件のダイアログボックスを開き、再度検索を行うことができます。

[削除]

一覧の ListBox のうちの選択した行を削除することができます。

[閉じる]

このウインドウを閉じます。

[一覧の ListBox]

一覧の ListBox は以下のようなセクションに分かれています。

・最初 ～ ----(対訳辞書追加)----

対訳辞書（english-to-japanese.lisp の分割辞書）の検索結果が表示されます。

・----(対訳辞書追加)---- ～ ----(意味ｺｰﾄﾞ辞書追加)----

意味コード辞書（ema-atr-sem-code.text の分割辞書）の検索結果が表示されます。

・----(意味ｺｰﾄﾞ辞書追加)---- ～ ----(形態素辞書追加)----

形態素辞書（ema-eng-morph-manual.dic の分割辞書）の検索結果が表示されます。

・----(形態素辞書追加)---- ～ 末尾

形態素辞書（ema-eng-morph.dic の分割辞書）の検索結果が表示されます。

・----(対訳辞書追加)----、----(意味ｺｰﾄﾞ辞書追加)----、----(形態素辞書追加)----

これを選んで[行編集]を押すと、それぞれの辞書に行単位に新規追加ができます。

辞書行編集のダイアログボックスが表示されます。

※制限事項：行編集や辞書登録をすると自動的に重複チェックが行われますが、下記のように自動的にコメントアウトや削除がなされた場合、一覧表示には反映されません。

※重複チェック：同じ語に異なる訳を登録しようとした場合は次のように処理されます。

(1) 置き換えていいかどうか尋ねます。

(2) 置換前の行がコメントアウトされ、置換後の行がその位置に追加されます。

(3) さらに、置換後の行がコメントアウトされた状態で存在する場合は削除されます。

3.6　辞書行編集

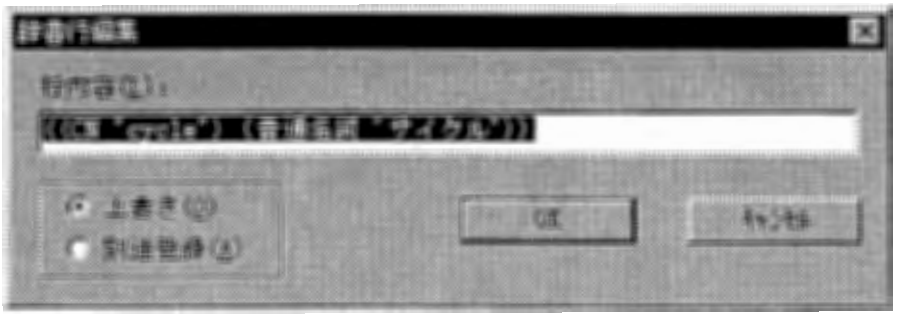

図 6　辞書行編集

辞書の内容を行単位に編集します。

[行内容]

行の内容を入力／編集してください。なお、[OK]を押す時に、書式が合っているかどうかのチェックが行われます。行の先頭に「;」を付けてコメントアウトするような変更（あるいはその逆）も可能です。

[上書き]

変更前の行を削除し、[行内容]と置き換えます。

手編集用でない方の形態素辞書（ema-eng-morph.dic の分割辞書）は上書きできません。

[別途登録]

変更前の行はそのまま残し、別途登録します。

なお、この場合でも重複や矛盾がないかどうかのチェックが行われます。

3.7　角川類語新辞典の検索

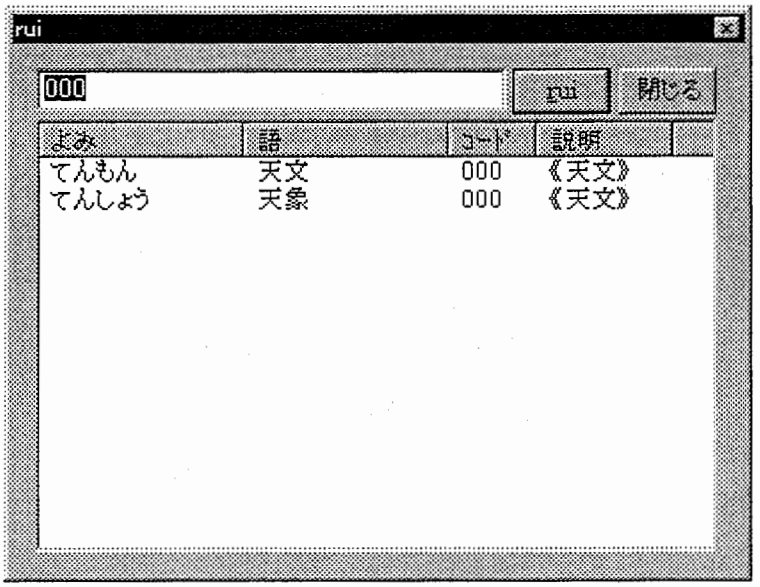

図7　ｒｕｉ

角川類語新辞典の検索を行います。

- 入力欄に検索する意味コード（半角の数字で）、または、よみ・語・説明（全角で）を入力し、[rui]を押すかリターンキーを押すと検索が実行されます。
- [よみ][語][コード][説明]を押すと、それぞれの欄の内容でソートして表示することができます。
- 検索結果のよみの部分をダブルクリックすると、辞書登録ダイアログボックスの[意味コード]の欄に意味コードの数字を挿入することができます。

3.8 分割辞書の再結合とUNIXへの転送

[凡例]

・作業用のテンポラリディレクトリを x:¥tmp と表す。

・?は?そのものを入力する。

・ディレクトリ名等は適宜読み替えてください。

(1) ファイルの結合 (DOS 窓を開いて)

(a)対訳辞書

```
copy /b C:¥DIV¥EJ¥?? x:¥tmp¥wk
echo (define-transfer-dic :e-j t>x:¥tmp¥EJ_SJIS
type x:¥tmp¥wk >> x:¥tmp¥EJ_SJIS
echo )>>x:¥tmp¥EJ_SJIS
del x:¥tmp¥wk
```

(b)意味コード辞書

```
copy /b C:¥DIV¥ES¥?? x:¥tmp¥ES_SJIS
```

(c)形態素辞書(-manual)

```
copy /b C:¥DIV¥EM¥?? x:¥tmp¥EM_SJIS
```

(2) ftp で x:¥tmp¥ES_SJIS 等を UNIX/Linux に転送 (ASCII モードで)

とりあえず /tmp あたりに転送する。

転送したら x:¥tmp¥ES_SJIS 等は削除してよい。

(3) mule-internal に変換 (UNIX/Linux で)

```
cd /tmp
coco ¥*sjis¥* ¥*internal¥* < EJ_SJIS > english-to-japanese.lisp
coco ¥*sjis¥* ¥*internal¥* < ES_SJIS > ema-atr-sem-code.text
coco ¥*sjis¥* ¥*internal¥* < EM_SJIS > ema-eng-morph-manual.dic
```

(4) 所定の位置にコピー (必要ならバックアップ付きで)

```
cp -ib -V numbered english-to-japanese.lisp /xxx/
cp -ib -V numbered ema-atr-sem-code.text /yyy/
cp -ib -V numbered ema-eng-morph-manual.dic /zzz/
```

※以上の内容を毎回入力するのが面倒であれば、適宜 .BAT やシェルを作っておくとよいでしょう。

第 4 章

応用プログラム

4.1　未知語を学習する仕掛け

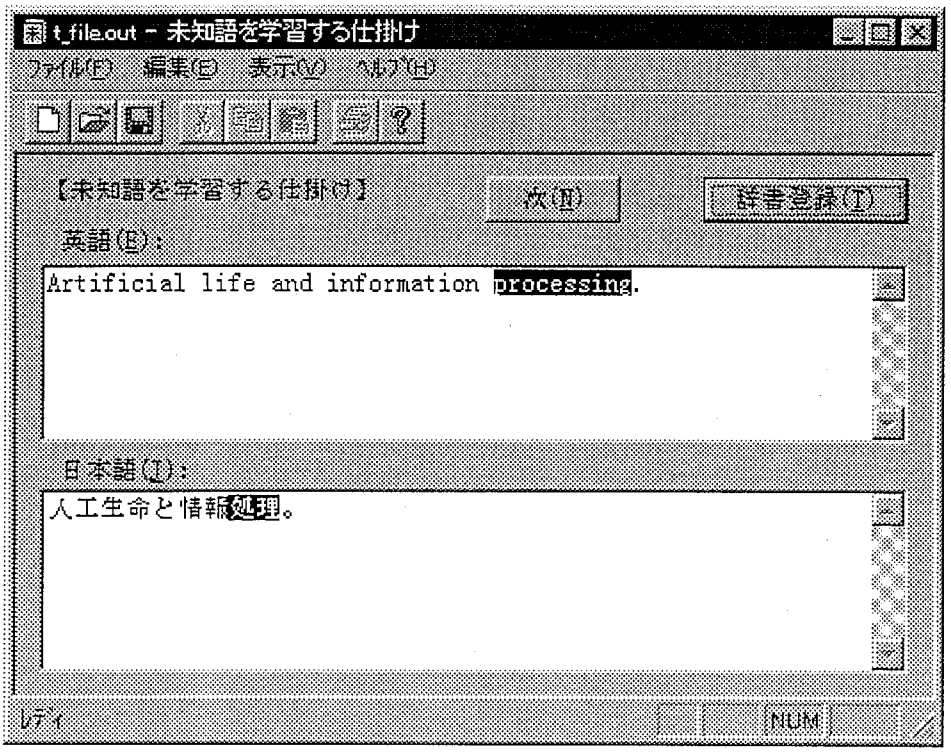

図 8　未知語を学習する仕掛け

辞書ツールの API を利用した応用プログラム例であり、翻訳結果を後編集し、未知語を辞書に登録するデモプログラムです。MICHIGO0.EXE を起動してください。

●使い方

(1) 英語の欄に英文を入力し、日本語の欄に翻訳結果を入力します。

（実際の運用時には[翻訳]ボタンを付け加え、それを押すことにより翻訳結果が表示されることを想定しています。）

(2) 翻訳結果には未知語がアルファベットの並びとして含まれていますので、後編集して修正します。

(3) 英語・日本語の辞書登録する範囲を選択し、[辞書登録]を押すと、選んだ範囲で辞書登録画面になります。

●翻訳結果のファイルを次々と読み込む機能

(1) メニューの [ファイル]-[開く] で次の形式のファイル（事前に Shift-JIS に変換してください）を開きます。

・emacs-lisp のツールの"T"での一括翻訳のファイル出力

```
source : "English"
target : "翻訳結果" (1.2345)
```

・make-result の出力

例：Lisp のプロンプトで

```
(make-result "/tmp/xxx" '(("JST" "open-syoroku")))
```

```
((E "English")
("翻訳結果" . 1.2345)
)
```

(2) 翻訳結果を後編集し、必要に応じて辞書登録します。

(3) [次] を押すと、後編集の結果がプログラム内部に記憶され、次の文へ移ることができます。

(4) 一連の後編集の結果は、メニューの [ファイル]-[上書き保存] または [名前を付けて保存] でファイルにセーブできます。

4.2 辞書分割ツール

1つの巨大な辞書を細かく分割するツールであり、「2.3 TDMT 辞書のセットアップ」の項目で使用したSPLDIC.EXEである。起動は Windows 95 の MS-DOS プロンプトを開いて行い、コマンドラインから以下のような引数と共に入力する。

[書式]

SPLDIC *mode* *辞書ファイル* *出力先*

[説明]

mode : ej、esem、emorph のいずれか

ej : 英日対訳辞書

esem : 英語意味コード辞書

emorph : 英語形態素辞書

辞書ファイル : 辞書ファイルの名前を指定する（必要ならパス名を付けて）。

出力先 : 分割辞書を出力するディレクトリ名。末尾に ¥ を付けること。

4.3 まとめて追加するツール

辞書の行の形式で書かれた追加リストを分割辞書にまとめて追加するツールである。対訳辞書・意味コード辞書・形態素辞書それぞれ独立して行う。追加する際には重複語のチェックがなされ、重複があった場合は置き換えるかどうかを尋ねる。起動は Windows 95 の MS-DOS プロンプトを開いて行い、コマンドラインから以下のような引数と共に入力する。

[書式]

・対訳辞書の場合

ADDLIST ej *fname* C:¥DIV¥EJ¥

・意味コード辞書の場合

ADDLIST　esem　*fname*　C:¥DIV¥ES¥

・形態素辞書の場合

ADDLIST　emorph *fname*　C:¥DIV¥EM¥　C:¥DIV¥EMA¥

[説明]

fname：追加リストのファイル名を指定する（必要ならパス名を付けて）。

C:¥DIV¥EJ¥等 ：分割辞書のディレクトリ名。末尾に ¥ が必要。

形態素辞書の C:¥DIV¥EMA¥：手編集用でない方の分割辞書

4.4　応用例：ディスクにすでにある辞書の重複語チェックの方法

以上のツールを使って、すでにディスク上にある辞書の重複語をチェックする方法を示す。以下、英日対訳辞書の場合の例で示す。

(1) 作業ディレクトリを新しく作る。

```
md x:¥tmp¥ej¥
```

(2) 空の分割辞書ファイルを作成。

```
spldic ej nul x:¥tmp¥ej¥
```

(3) チェック対象の分割辞書を結合。

分割辞書毎に並べ替えた状態で結合することになるが、これが速度の面で重要。

```
copy /b C:¥DIV¥EJ¥?? x:¥tmp¥ALL
```

（?は?そのものを入力する）

(4) 空のファイルに対して重複チェックしながら追加。

結合ファイル読み込み時にしばらく静かになるが、ハングアップしたわけではないので心配しないように。

ADDLIST ej x:¥tmp¥ALL x:¥tmp¥ej¥

(5) 必要なら元の場所にエクスプローラーでコピーする。

【注意】copy コマンドはサイズ０のファイルのコピーを省いてしまうので、エクスプローラー等で行うこと。

参考文献

(1) 古瀬 蔵, 隅田英一郎, 飯田 仁: “経験的知識を活用する変換主導型機械翻訳”, 情報処理学会論文誌, Vol.35, No.3, pp.414-425(1994)

(2) 河井 淳: “TDMT 翻訳知識作成ツールの利用方法”, ATR テクニカルレポート TR-IT-0088, (1994-12)

(3) 大野晋, 浜西正人: “角川類語新辞典 CD-ROM 版”, 角川書店, (1989)

(4) 鷹尾 和享, 柏岡 秀紀, 白井 諭: “TDMT 辞書ツール説明書(英日版・開発者向け)”, ATR テクニカルレポート TR-IT-0314, (1999-9)

(5) 鷹尾 和享, 柏岡 秀紀, 白井 諭: “異なる辞書を利用した意味コードの自動付与”, 情報処理学会第 59 回全国大会, 1N-07, (1999-9)