第３章

辞書構築手順

EDBroW 利用にあたっては，最初に CD-ROM で用意されたテキスト形式の辞書（以下，テキスト辞書と呼ぶ）を，EDBroW から利用できる形式の辞書（以下，システム辞書と呼ぶ）に変換します．変換ツールである辞書トランスレータは，UNIX 上，もしくは Windows 上で動作します．ツールは，EXEC 以下にあります．dtrans.exe が，パソコン用の辞書トランスレータ，dtrans が Sun SparcStation 用の辞書トランスレータです．

3.1 辞書構築の概要

3.1.1 テキスト辞書のマージについて

提供されるテキスト辞書は，格納されているデータにより，単語辞書，対訳辞書，共起辞書，概念辞書，日本語コーパスに大別されます．概念辞書はさらに，概念体系辞書，概念記述辞書，概念見出し辞書の３種類に分かれています．

EDBroW(V1.3)から日本語コーパスの検索・表示機能をサポートするために辞書構築処理においても日本語コーパスのCD-ROMデータからの変換を行なえるようになっております．システム辞書の内部形式はまったく別のものですが，検索・表示システムや辞書構築手順のユーザインタフェースには違いはありません．

概念辞書を除く各辞書は，その言語識別（日本語／英語，対訳の場合は日英／英日，日本語コーパス）に応じて別ファイルとなります．また，専門用語（情報処理）の辞書も別ファイルとなっています．したがって，単語辞書であれば，最大で，日本語単語辞書，英語単語辞書，日本語専門用語単語辞書，英語専門用語単語辞書，の４種類となるわけです．ユーザによっては，日本語単語辞書だけを購入した，という場合もあるでしょう．

EDBroW では，一度に扱えるシステム辞書は種類ごとに１つです．すなわち，単語辞書が１つ，対訳辞書が１つという形です．このため，辞書トランスレータでは，複数ある同種のテキスト辞書をマージして，一つのシステム辞書として作成する機能を有しています．

概念辞書を除くシステム辞書は，言語識別，基本語／専門用語の別を格納するように設計されており，辞書をマージした場合でも情報は失われません．また，ブラウザでそれらの種別を指定して検索することも可能です．

※ 概念見出し辞書，概念体系辞書は，基本語／専門用語の別に応じて，２種類のテキスト辞書が提供されますが，この種別を格納するように設計されてはいません．これは，基本語／専門用語で概念識別子が重複することはないため，ユーザの混乱は招かない，という理由によります．

辞書をマージして作成するか，テキスト辞書ごとに別の辞書として作成するか，はユーザの自由です．EDBroW では，起動時に初期化ファイルに記述されたデフォルトのシステム辞書を開くようになっていますが，起動後にシステム辞書を切り換えることもできます．

辞書をマージした場合，例えば単語辞書を概念識別子で検索することにより，その概念につながる日本語／英語双方の単語を一度に表示することができるようになります（対訳辞書の場合も同様）．逆に，辞書をマージすると，辞書サイズが大きくなる分，検索速度が若干低下しますが，さほど問題はありません．

3.1.2 辞書構築の流れ

デフォルトでは，システム辞書は辞書トランスレータを実行しているカレントディレクトリに作成されます（ただし，オプションなしの起動では，パス名を直接指定することが可能）．また，作業領域も，デフォルトではカレントディレクトリですが，環境変数により，別のディレクトリを指定することもできます．指定する環境変数は，DTRANS_TMP です．

UNIX の場合は，コマンドラインから，

```
% setenv DTRANS_TMP /work
```

などと指定するか，.cshrc などで指定して下さい．

必要な作業領域は，辞書の種類にもよりますが，最大でも作成されるシステム辞書サイズの 2 倍以内です．また，システム辞書サイズは，最大でも元となるテキスト辞書サイズを越えることはありません．

※ 概念見出し辞書の場合のみ，システム辞書サイズはテキスト辞書サイズ（複数ある場合は合計）の 75% 程度になりますが，それ以外の辞書では，40% 〜 60% 程度になります．

辞書の構築手順は，おおまかに言って以下のようになっています．

- ♦ 最初に作成するシステム辞書の種類を指定する．
- ♦ システム辞書の種類に対応するテキスト辞書を，必要な数だけ指定する．
- ♦ テキスト辞書の読み込みが完了すると，検索のためのインデックスなどを作成し，システム辞書が完成する．

共起辞書のみ，他の辞書と構築手順が若干違います．これについては，3.4.1 を参照して下さい．

次節に，単語辞書を例にとって，システム辞書構築手順を詳説します．

3.2 Windows 上でのシステム辞書構築

Windows 上の辞書トランスレータは，インタラクティブに動作します．辞書の種類，パス名などを指定するプロンプトが表示されますので，必要な情報を入力していって下さい．

なお，途中で終了する場合には，q を入力して下さい．処理の途中で，Ctrl+C や［ファイル］－［終了］コマンドで終了すると，作業ファイルや作成途中のシステム辞書ファイルが消去されずに残ってしまうことがあります．

※途中で中断した場合には，fdc?????.tmp, idx?????.tm0, idx?????.tm1, srt?????.tmp などのファイル（????? はプロセスID）が残ることがあります．これらはかなり大きなファイルとなりますので，消去して下さい．

それでは，実際に単語辞書を作成してみます．

まず Windows を起動し，dtrans のアイコンをダブルクリックして下さい（または，ファイルマネージャから dtrans.exeをダブルクリック）．辞書トランスレータが起動し，以下のようなメッセージが表示されます．

```
EDBroW 辞書トランスレータ　Ver.1.3
Copyright (c) JAPAN ELECTRONIC DICTIONARY RESEARCH INSTITUTE, LED 1998
***** 辞書種類の指定 *****
  1. 単語辞書　2. 対訳辞書　3. 共起辞書
  4. 概念見出し辞書　5. 概念体系辞書 6. 概念記述辞書
  7. コーパス辞書
辞書の種類を指定して下さい (q:終了)
[入力] >>
```

最初に，

```
辞書の種類を指定して下さい（q：終了）
[入力] >>
```

のプロンプトで，どのシステム辞書を作成するかを選択します．作成したい辞書の番号を入力し，リターンキーを押して下さい．ここでは，単語辞書作成なので，1 を入力します．

続いて，

```
作成する辞書のパス名を入力して下さい
（デフォルト：カレントディレクトリに wd_sys.dic を作成）
[入力] >>
```

というプロンプトが表示されます．リターンのみを入力すると，カレントディレクトリに，wd_sys.dic というシステム辞書ファイルが作成されますが，パス名を指定すると，そのパス名でシステム辞書を作成します．パスの指定が不正である場合には，再度上記プロンプトが表示されますので，正しいパス名を入力して下さい．なお，辞書ファイルのデフォルト名は，辞書の種類により異ります（→3.3）．

また，すでに同じパス名のファイルがある場合，ファイルを削除してよいかどうかの確認を求めます．ここで y を入力すると，既存のファイルは削除され，新たにシステム辞書ファイルを作成します．n を入力した場合には，システムは処理を終了します．

次に，

```
テキスト辞書のパス名を入力して下さい (q:指定終了)
[入力] >>
```

というプロンプトが表示されます．単語辞書の CDROM をセットし，テキスト辞書（CDROM 上）のパス名を入力してリターンキーを押して下さい．テキスト辞書の読み込みを開始します．テキスト辞書の読み込みにかかる時間は，作成する辞書の種類や，計算機のパフォーマンスにより異なります．

なお，指定したパスに誤りがある場合や，別の種類のテキスト辞書を指定した場合にはエラーが表示されますので，入力をやり直してください

テキスト辞書の読み込みが完了すると，再び上記のプロンプトが表示されます．別の単語辞書をマージする場合には，CDROM を入れかえ，再度テキスト辞書のパス名を入力して下さい．マージする必要がない場合，あるいはマージするすべてのテキスト辞書の読み込みが完了した場合には，q を入力します．

テキスト辞書の読み込みが完了すると，自動的にシステム辞書の構築にに移ります．ここで，読み込みに要するのと同じくらいの時間がかかります．

辞書の構築が終了すると，ウィンドウ上に以下のようなメッセージが表示されます．

******* 単語辞書の作成完了 *******
変換レコード数　390164 個
異常レコード数　　0 個

これで，辞書の変換は完了です．処理が終了しても，ウィンドウは消去されませんので，［ファイル］－［終了］コマンドを選択してウィンドウを消去して下さい．

以下に，４種類の単語辞書をマージして，システム辞書（単語）を作成する場合の表示と入力の例を示します（CDROM ドライブのデバイスを E: ，作成する辞書のパス名を D:\DICT\WD_SYS.DIC と仮定，下線部分が入力を示す）．

```
***** 辞書種類の指定 *****
  1. 単語辞書  2. 対訳辞書  3. 共起辞書
  4. 概念見出し辞書  5. 概念体系辞書 6. 概念記述辞書
  7. コーパス辞書
辞書の種類を指定して下さい (q:終了)
[入力] >> 1                          ←辞書種類
***** 単語辞書の作成開始 *****
作成する辞書のパス名を入力して下さい
（デフォルト：カレントディレクトリに wd_sys.dic を作成）
[入力] >> D:\DICT\WD_SYS.DIC         ←システム辞書パス名
ファイル D:\DICT\WD_SYS.DIC を消去しても良いですか [Y/N] y
システム辞書 D:\DICT\WD_SYS.DIC を作成します

テキスト辞書のパス名を入力して下さい (q:指定終了)
[入力] >> E:\JWD.DIC                 ←テキスト辞書パス名
テキスト辞書のパス名を入力して下さい (q:指定終了)
[入力] >> E:\EWD.DIC                 ←テキスト辞書パス名
テキスト辞書のパス名を入力して下さい (q:指定終了)
[入力] >> C:\TJWD.DIC                ←テキスト辞書パス名（誤り）

[ERROR!] ファイルがオープンできません : C:\TJWD.DIC

テキスト辞書のパス名を入力して下さい (q:指定終了)
```

```
[入力] >> E:\TJWD.DIC                    ←テキスト辞書パス名
テキスト辞書のパス名を入力して下さい (q:指定終了)
[入力] >> E:\TEWD.DIC                    ←テキスト辞書パス名
テキスト辞書のパス名を入力して下さい (q:指定終了)
[入力] >> q                              ←指定終了コマンド
データ読み込み完了

固定長パートの書き込み中...終了
インデックス(#1)の作成開始
  ソート中...終了
  書き込み中...終了
インデックス(#2)の作成開始
  ソート中...終了
  書き込み中...終了
インデックス(#3)の作成開始
  ソート中...終了
  書き込み中...終了

******* 単語辞書の作成完了 *******
  変換レコード数    927658 個
  異常レコード数         0 個
```

対訳辞書，概念辞書の場合も，作成方法はまったく同じです．ただし，インデックスを３種類作成するのは単語辞書だけです．

概念辞書の場合，一つの CD-ROM に複数の辞書（概念体系，概念記述，概念見出し）が入っていますので，パス名に注意して下さい．専門用語の概念辞書（概念体系，概念見出しのみ）をマージする場合の方法は，単語辞書と同様です．

共起辞書の場合には，基本的には上記と同じ手順ですが，いくつか情報を追加する必要があります．詳しくは，3.4.1 に記します．

コーパスのシステム辞書を作成する際には，ＳＪＩＳ版ＣＤ－ＲＯＭをセットしてください．

3.3 UNIX 上でのシステム辞書構築

UNIX 版では，コマンドラインオプションを指定することができます．オプションを指定しない場合はインタラクティブに動作し，辞書の指定方法などはパソコン版とまったく同じですので，省略します．但し，UNIX版にはコーパスの変換機能はありません．これは，UNIX用に辞書の管理システムが公開されているためで，ＥＤＲのホープページを参照ください．

コマンドラインオプションでは，辞書種別を指定します．この場合，テキスト辞書は stdin から入力します．

オプションと，作成されるシステム辞書の関係は以下のようになっています．

-WD　単語辞書
-BI　対訳辞書
-CC　共起辞書
-CI　概念見出し辞書
-CH　概念体系辞書

-CR　概念記述辞書

同種別ファイルのテキスト辞書をマージする場合には，

```
% cat JWD.TAB EWD.TAB | dtrans -WD
```

のようにして起動して下さい．後は，自動的にシステム辞書が作成されます．なお，コマンドラインオプションを指定して起動した場合には，辞書はカレントディレクトリに以下のファイル名で作成されます（これらは，Windows 版／UNIX版のオプションなしによる起動時 のデフォルトファイル名でもあります）．

単語辞書	wd_sys.dic
対訳辞書	bi_sys.dic
共起辞書	cc_sys.dic
概念体系辞書	ch_sys.dic
概念記述辞書	cr_sys.dic
概念見出し辞書	ci_sys.dic

すでに同じ名前のファイルがあった場合には，辞書トランスレータは処理を中断します．

注：UNIX版にはコーパスの変換処理はありません。

3.4　注意点

3.4.1　共起辞書の構築手順

共起辞書には，概念識別子の代わりに「補足付き概念説明」と呼ばれるデータが入っているレコードが多くあります．このデータを通常の概念識別子と同等に扱うためには，共起辞書構築時に，他の辞書とは異なる操作が必要です．すなわち，補足付き概念説明に対して概念識別子を自動付与し，概念識別子と補足付き概念説明文字列との対応を，概念見出し辞書のテキスト形式と同じフォーマットで格納することにより，補足付き概念説明を通常の概念識別子と同じ枠組みで利用しようというもので，これを実現するため若干の操作が必要となるわけです．以下に，その手順を説明します．

辞書トランスレータを起動し，辞書種類として 3（共起辞書）を入力すると，最初に以下のプロンプトが表示されます．

補足付概念説明の格納方法を指定して下さい
1. 格納しない（0x0L として格納）
2. 重複を許して格納（識別子は自動付与）
3. 重複を除いて格納（識別子は自動付与）

[入力] >>

ここで，格納方式を指定します．各方式の意味は以下の通りです．

格納しない：

補足付き概念説明が概念識別子の代わりにはいっている場合，概念識別子を 0 として格納します．したがって，補足付き概念説明の持つ情報は失われ，意味情報が不完全になりますが，概念見出し辞書に対する変更はありません．

重複を許して格納：

補足付き概念説明があった場合，他の辞書の概念識別子とぶつからない範囲の概念識別子を自動生成し，これを格納します．テキスト辞書中に同じ補足付き概念説明が現れた場合でも，別の概念識別子を付与します．概念識別子は，補足付き概念説明がテキスト辞書中に出現した順に，c00001 から一つづつ付与されます．

重複を除いて格納：

補足付き概念説明があった場合，他の辞書の概念識別子とぶつからない範囲の概念識別子を自動生成し，これを格納します．テキスト辞書中に同じ補足付き概念説明が現れた場合には，同じ概念識別子を付与します．概念識別子は，補足付き概念説明がテキスト辞書中に出現した順（同じものなら最初に出現したもの）に，c00001 から一つづつ付与されます．

1 を選択した場合には，その他の辞書変換の場合とと同じ処理に戻ります． 2 または 3 を選択した場合，以下のプロンプトが表示されます．

```
補足付概念説明ファイルのパス名を入力して下さい
（リターンのみ：カレントディレクトリの cctoci.txt に格納）
[入力] >>
```

ここで，パス名を入力します．このファイルは，辞書トランスレータが自動付与した概念識別子と，補足付き概念説明文字列との対応をとるもので，実際には概念見出し辞書（テキスト）と同じ形式で，概念識別子のフィールドに自動生成した識別子を，日本語概念説明のフィールドに補足付き概念説明文字列を格納します．

ファイル名入力が終われば，その後は他の辞書構築手順と同様の処理に戻ります．

ここで作成された，補足付き概念説明格納ファイルは，概念見出し辞書を作成する際にマージして下さい．補足付き概念説明が，他の概念識別子とほとんど同様に利用できるようになります．ただし，c00001以降の概念識別子は共起辞書のみに有効であり，これを使って他の辞書を検索することはできませんので，ご注意下さい．

共起辞書構築時の注意点

1 共起辞書では，システム辞書サイズ縮小のために，変換時にかなりのメモリを必要とします．補足付き概念説明の格納方法として，1 または 2 を選択した場合で，パソコン本体のメモリが 8MB 程度の場合，変換はできますが相当の時間がかかりますのでご注意下さい．

2 補足付き概念説明の格納方法として，重複を許す／重複を除くのどちらを選択するか，は作成されるファイルサイズとシステムのメモリ容量とのトレードオフとなります．重複を許す場合，メモリは 8MB 程度でも処理できますが，作成されるテキストファイルの大きさは倍以上になります（日本語共起辞書のみの場合で 20MB 程度）．逆に，重複を許す場合，メモリは少なくとも 20MB は必要ですが，テキストファイルののサイズは半分以下になります（日本語共起辞書の場合で 9MB 程度）．

3 UNIX 版トランスレータで，オプションつき（-CC）で起動した場合，補足付き概念説明の格納方式は，自動的に 3（重複を除いて格納）になり，データはカレントディレクトリの cctoci.txt に格納されます．

3.4.2 概念見出し辞書の構築

EDBroW では，単語辞書，対訳辞書，共起辞書には概念識別子のみを格納し，概念識別子の説明であるところの概念見出しは，概念見出し辞書として別に格納しています．このため，概念見出し辞書がないと，各辞書レコードの語義が表示されません．

したがいまして，本システムの利用にあたっては，なるべく概念辞書一式を購入されることをお勧めします．

ただし，事情により，概念辞書一式が購入できない場合もあるかと思います．辞書トランスレータのオプションとして，mkcitxt というユーティリティを用意してあります．これは，単語／対訳辞書の中から，概念見出しに相当するエントリのみを抜き出し，テキスト形式の概念見出し辞書と同等のフォーマットで出力するツールです．このツールにより作成されたテキストファイルを，辞書トランスレータで概念見出し辞書に変換することができます．

詳しくは，dtrans ディレクトリの中にある，util ディレクトリ以下の，readme.{euc, sj} を参照して下さい．

なお，共起辞書の変換中に「補足付き概念説明」を格納するファイルを作成した場合には，このファイルもマージして，概念見出し辞書を作成して下さい．