第 7 章

ファイルフォーマット

7.1 システム辞書

7.1.1 基本構造

システム辞書とは，本システムが直接参照する辞書であり，以下に示す 6 種類がある．

単語辞書
対訳辞書
共起辞書
概念体系辞書
概念記述辞書
概念見出し辞書

システム辞書は読み出し専用とし，EDBroW の動作によってシステム辞書を書き換えることはない．また，単語 / 対訳 / 共起辞書については，言語識別，基本語 / 専門用語の区別によらず，同一ファイルとすることが可能である*．

各システム辞書は，それぞれ1つのファイルとして作成される．辞書ファイルは，以下の6つのパートにより構成される．

辞書情報パート
可変長辞書パート
固定長辞書パート
インデックスパート #1 ~ #3

インデックスパートは，辞書を検索するためのインデックス部分であり，一つの辞書に対して3つまでのインデックスパートを持つことができる．辞書情報パートには，辞書識別子，各辞書 / インデックスパートのレコードサイズ，先頭オフセットなどが格納される．

図7.1に，例として単語辞書の構造を示す．

* ただし，共起辞書の専門用語辞書，および日本語共起パータン副辞書については，基本語辞書と形式がまったく異なるため，本システムではサポートしない．

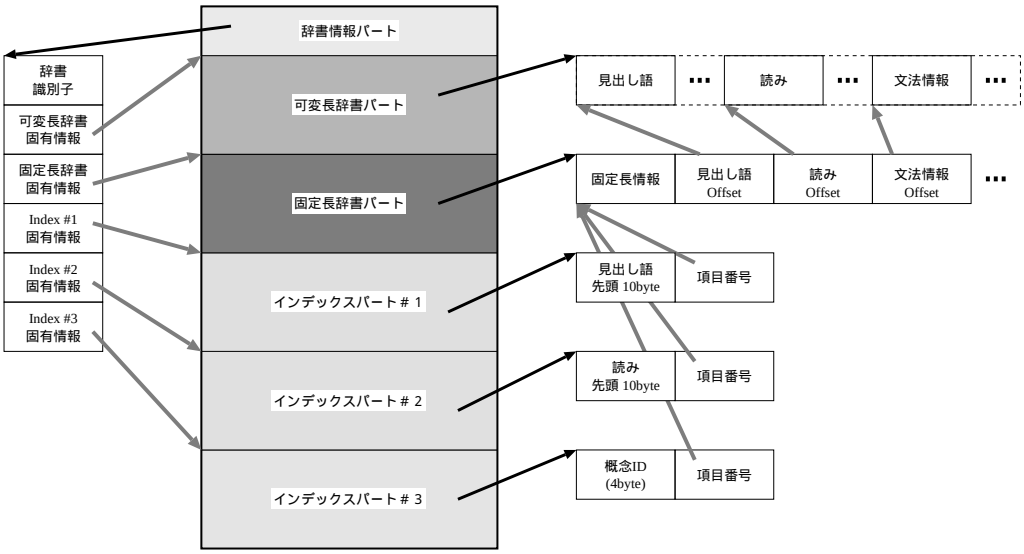

図 7.1 単語辞書の構造

辞書情報パート

フィールド名	格納データ	データ型	データサイズ
辞書固有情報	辞書識別子	char	1byte
	項目数	long	4byte
可変長パート	先頭オフセット	long	4byte
固定長パート	先頭オフセット	long	4byte
固有情報	レコードサイズ	char	1byte
	固定長情報サイズ	char	1byte
インデックス	先頭オフセット	long	4byte
#1 ~ #3	インデックスサイズ	char	1byte
固有情報	インデックスタイプ	char	1byte
	データフラグ	char	1byte
	キー格納位置	char	1byte

表 7.1 辞書情報パートのレコード形式

表 7.1 に，辞書情報パートのレコード形式を示す．

ここで，各インデックス固有情報における，"データフラグ"とは，そのインデックスパートに，辞書情報がそのまま格納されている（固定長辞書のかわりになる）ことを示し，辞書情報が格納されている場合は 1 ，そうでない場合には 0 を格納する* ．また，"インデックスタイプ"とは，格納されているキーが固定長か可変長（の固定長先頭バイト）か，を示すものであり，可変長の場合には 1 を，固定長の場合には 0 を格納する．"キー格納位置" は，固定長辞書中で，当該キー（可変長キーの場合は先頭からのオフセット）が格納されている位置を示す．

* これは，検索速度の向上と，辞書サイズの圧縮を意図したものであり，システム辞書では，特に頻繁に検索される概念見出し辞書と，概念体系辞書がこの形式で作成される．

辞書識別子は，表 7.2 に示すコードを格納する．

システム辞書名	コード	略称
単語辞書	1	WD
概念体系辞書	2	CH
概念記述辞書	3	CR
概念見出し辞書	4	CI
対訳辞書	5	BI
共起辞書	6	CC

表 7.2　辞書識別子

表 7.3に，各システム辞書の各属性を示す．表中，各インデックスパートにおいて，"○ " は可変長キー，"☆" は固定長キーの，それぞれインデックスとなること (インデックスタイプに対応) を示し，"★" はさらに，固定長辞書の役割を果すこと (データフラグに対応) を示す．また，固定長辞書 / インデックスの，記号横の ( ) 内の数字は，そのレコードのデータサイズ (byte) を示す．

辞書名	識別子	可変長パート	固定長パート	Index #1	Index #2	Index #3
単語辞書	WD	○	○(27)	○(14)	○(14)	☆(8)
概念体系辞書	CH	−	−	★(8)	★(8)	−
概念記述辞書	CR	−	○(14)	☆(8)	☆(8)	☆(6)
概念見出し辞書	CI	○	−	★(8)	−	−
対訳辞書	BI	○	○(21)	○(14)	☆(8)	−
共起辞書	CC	○	○(47)	○(14)	○(14)	○(14)

表 7.3 各辞書情報のサイズ

以下に，各辞書ごとに，その詳細な構造を示す．

7.1.2　単語辞書

表 7.4に，単語辞書のレコード形式を示す．

可変長のフィールドについては，可変長辞書パートに格納し，固定長辞書パートには，そのオフセット，すなわち辞書の先頭からの可変長データ格納位置のバイト数を格納する．また，可変長辞書パート中では，同一の文字列が頻繁に格納されることのないよう，辞書トランスレータにおいて圧縮処理を行う．なお，テキスト辞書中のEUC コードは，Shift-JIS コードに変換して格納する*．

可変長フィールドのうち，「発音情報 / 構成語情報」のフィールドは，当該辞書レコードの種別によって，格納される情報が異なる．すなわち，慣用句レコードの場合 (慣用句フラグが1である場合) には，構成語情報が，そうでない場合には，発音情報が格納される．

※ここに示した，可変長フィールドの可変長辞書パートへの格納方法は，以下に示すすべてのシステム辞書に共通である．

フィールド名	データ型	データサイズ	検索キー	備考
単語見出し	文字列	可変長	○	Index #1
読み	文字列	可変長	○	Index #2
発音情報 / 構成語情報	文字列	可変長	×	
概念識別子	unsigned long	4byte	○	Index #3
固定長情報	ビットフィールド	3byte		
左連接情報		8ビット	×	
右連接情報		8ビット	×	
品詞情報		6ビット	△	
言語識別		1ビット	△	
レコード種別		1ビット	△	
慣用句フラグ		1ビット	×	
可変長文法情報	文字列	可変長	×	
構文木	文字列	可変長	×	

表 7.4　単語辞書のレコード形式

固定長情報中の，左連接情報・右連接情報・品詞は，コード数値 (→ 5.8) により格納する．言語識別は，日本語の場合0を，英語の場合1を格納し，慣用句フラグは，当該レコードが慣用句である場合に1を，そうでない場合には0を格納する．

「検索キー」の欄において，"○ " は，システム辞書の検索において検索キーとなることを示す．検索キーとなるフィールドに対しては，これによって辞書を昇順にソートしたインデックスパートを作成し，検索を高速に行えるようにする．また，"× " は検索キーとはならないことを示し，"△ " は，直接の検索キーにはならないが，フィルタキーとして利用可能であることを示す．この表記は，以下に述べる辞書の構造すべてに共通である．

各インデックスパートのレコード形式を 表 7.5 に示す．

インデックス	フィールド名	データ型	データサイズ	備考
#1	単語見出し	文字列	10byte	単語見出しの先頭 10byteを格納
	項目番号	unsigned long	4byte	負でない整数
#2	読み	文字列	10byte	読みの先頭 10byteを格納
	項目番号	unsigned long	4byte	負でない整数
#3	概念識別子	unsigned long	4byte	概念識別子を格納
	項目番号	unsigned long	4byte	負でない整数

表 7.5　単語辞書のインデックスパート レコード形式

単語見出し / 読みは，単語辞書の単語見出し / 読みフィールドの先頭10バイトと等しい．長さが10バイト未満の場合は，'¥0'で終端する．10バイト以上の場合は，終端しない．また，概念識別子は単語辞書の概念識別子と等しく固定長である．

項目番号は，単語辞書の各エントリーに付与した非負の整数である．項目番号に固定長辞書パートのレコードサイズを掛けたものに，固定長辞書パートの先頭オフセットを加えると，そのエントリーのオフセットとなる．

インデックスパートは，キーとなるシステム辞書のフィールドに対して昇順にソートする．すなわち，単語見出し / 読みの10バイトを超える部分に対しても，昇順であることを保証する．ここでいう昇順とは，単純に文字コードの比較によるものであり，テキスト辞書の順番は考慮しなくて良い．また，インデックスパートのレコードは，単語辞書のレコードと1対1に対応するため，それらの総数は等しい．

7.1.3 対訳辞書

表 7.6 に対訳辞書のレコード形式を示す．

フィールド名	データ型	データサイズ	検索キー	備考
単語見出し	文字列	可変長	○	Index #1
概念識別子	unsigned long	固定長	○	Index #2
固定長情報	ビットフィールド	1byte		
品詞		6ビット	△	
レコード種別		1ビット	×	
言語対識別		1ビット	△	
対訳情報	文字列	可変長	×	

表 7.6　対訳辞書のレコード形式

言語対識別とは，日本語→英語の対訳レコードであるか，英語→日本語であるか，を示すものであり．日英の場合は 0 を，英日の場合は 1 を格納する．また，レコード種別は，当該レコードが専門用語であるかどうか，を格納するものであり，値としては 1/0をとる．

訳語情報は，テキスト形式のテキスト辞書に格納されている文字列を，そのまま格納する．

インデックスパートのレコード形式を，表 7.7 に示す．

インデックス	フィールド名	データ型	データサイズ	備考
#1	単語見出し	文字列	10byte	単語見出しの先頭10byteを格納
	項目番号	unsigned long	4byte	負でない整数
#2	概念識別子	unsigned long	4byte	概念識別子を格納
	項目番号	unsigned long	4byte	負でない整数

表 7.7　対訳辞書のインデックスパート レコード形式

7.1.4 共起辞書

表 7.8 に共起辞書のレコード形式を示す．

言語識別は，日本語の場合 0 を，英語の場合 1 を格納する．品詞，関係子については，5.8 に示すコード番号を格納する．

表記順とは，句見出しと受け側 / 係り側単語との順序の対応をとるもので，句見出し中で係り側→受け側の順に並ぶ場合には0 を，逆の場合には 1 を格納する．

関係子方向とは，表層の係り受け関係と，意味情報における関係との対応をとるもので，受け側概念識別子が意味情報中の始点概念となる場合に 0 を，そうでない場合に 1 を格納する．

フィールド名	データ型	データサイズ	検索キー	備考
受け側単語見出し	文字列	可変長	○	Index #1
受け側補助情報	文字列	可変長	×	
係り側単語見出し	文字列	可変長	○	Index #2
係り側補助情報	文字列	可変長	×	
共起関係子	文字列	可変長	△	
関係単語表記	文字列	可変長	×	
受け側概念識別子	unsigned long	4byte	△	
係り側概念識別子	unsigned long	4byte	△	
固定長情報	ビットフィールド	3byte	×	
言語識別		1bit		
係り側品詞		4bit		
受け側品詞		4bit		
概念関係子		8bit		
概念関係子方向		1bit		
表記順		1bit		
受け側慣用句フラグ		1bit		
係り側慣用句フラグ		1bit		
表層共起頻度	int	2byte	×	
共起項目頻度	int	2byte	×	
受け側共起要素頻度	int	2byte	×	
係り側共起要素頻度	int	2byte	×	
例文	文字列	可変長	×	

表 7.8　共起辞書のレコード形式

その他フラグに関しては，コードもしくは 1/0に変換し，ビットフィールドとして格納する．

なお，例文は，テキスト辞書では複数例文が格納されているが，システム辞書では，'|'で区切った形式により，最大 5 例文を格納する．

インデックスパートのレコード形式を，表 7.9 に示す．

インデックス	フィールド名	データ型	データサイズ	備考
#1	受け側単語見出し	文字列	10byte	受け側単語見出しの先頭10byteを格納
	項目番号	unsigned long	4byte	負でない整数
#2	係り側単語見出し	文字列	10byte	係り側単語見出しの先頭10byteを格納
	項目番号	unsigned long	4byte	負でない整数

表 7.9　共起辞書のインデックスパート レコード形式

7.1.5　概念体系辞書

表 7.10 に概念体系辞書のレコード形式を示す．

この形式のレコードは，以下のような概念体系のデータを表現したものである．

概念識別子1(下位概念) − kind-of→ 概念識別子2(上位概念)

フィールド名	データ型	データサイズ	検索キー	備考
概念識別子1	unsigned long	4byte	○	Index #1
概念識別子2	unsigned long	4byte	○	Index #2

表 7.10　概念体系辞書のレコード形式

インデックスパートのレコード形式を，表 7.11に示す．概念体系辞書では，検索の高速化と，辞書ファイルのサイズ縮小のために，インデックスパートが固定長辞書の役割を果たす．

インデックス	フィールド名	データ型	データサイズ	備考
#1	概念識別子1	unsigned long	4byte	概念識別子1を格納
	概念識別子2	unsigned long	4byte	概念識別子2を格納
#2	概念識別子2	unsigned long	4byte	概念識別子2を格納
	概念識別子1	unsigned long	4byte	概念識別子1を格納

表 7.11　概念体系辞書のインデックスパート レコード形式

7.1.6　概念記述辞書

表 7.12 に，概念記述辞書のレコード形式を示す．

フィールド名	データ型	データサイズ	検索キー	備考
概念識別子1	unsigned long	4byte	○	Index #1
概念識別子2	unsigned long	4byte	○	Index #2
概念関係子	char	1byte	△	
真偽値	ビットフィールド	1bit	△	
記述区分	ビットフィールド	1bit	△	

表 7.12　概念記述辞書のレコード形式

この形式のレコードは，以下のような概念記述のタプルを表現したものである．

概念識別子1 (始点概念) − 概念関係子 / 真偽値 → 概念識別子2 (終点概念)

記述区分とは，当該概念記述が，実例としてEDR コーパス中に出現したもの(E記述)であるか，これに対して概念体系上で抽象化を行なったもの(I記述)であるかを示し，E記述の場合 0 を，I記述の場合 1 を格納する．

インデックスパートのレコード形式を，表 7.13に示す．

インデックス	フィールド名	データ型	データサイズ	備考
#1	概念識別子1	unsigned long	4byte	概念識別子1を格納
	項目番号	unsigned long	4byte	負でない整数
#2	概念識別子2	unsigned long	4byte	概念識別子2を格納
	項目番号	unsigned long	4byte	負でない整数

表 7.13　概念記述辞書のインデックスパート レコード形式

7.1.7 概念見出し辞書

表 7.14に，概念見出し辞書のレコード形式を示す．

フィールド名	データ型	データサイズ	検索キー	備考
概念識別子	unsigned long	4byte	○	Index #1
日本語概念見出し	文字列	可変長	×	
英語概念見出し	文字列	可変長	×	
日本語概念説明	文字列	可変長	×	
英語概念説明	文字列	可変長	×	

表 7.14 概念見出し辞書のレコード形式

可変長辞書パートでは，各概念識別子 に対応する，日本語概念見出し，日本語概念説明，英語概念見出し，英語概念説明を，TAB で区切った形式以下の形式で格納し，NULL 文字により終端する．

〈日本語概念見出し〉¥t〈日本語概念説明〉¥t〈英語概念見出し〉¥t〈英語概念説明〉

インデックスパートのレコード形式を表 7.15に示す．概念見出し辞書では，インデクスパートが固定長辞書としての役割も果たす．

インデックス	フィールド名	データ型	データサイズ	備考
#1	概念識別子 オフセット	unsigned long unsigned long	4byte 4byte	概念識別子を格納 可変長辞書パートの オフセットを格納

表 7.15 概念見出し辞書のインデックスパート レコード形式

7.2 ユーザ情報ファイル

7.2.1 基本構造

ユーザ情報ファイルは，レコード・バッファ内のユーザ情報を1レコードとし，これをテキスト形式で格納しているファイルである．

ユーザ情報ファイル一つに格納されるユーザ情報は一つである．また，ユーザ情報ファイルであることを明示するため，ファイルの拡張子は .uif に統一する．

ユーザ情報ファイルの形式を以下に示す．

〈ユーザ情報ファイル〉::=〈キーワード〉¥n〈ユーザ情報レコード〉。。。

〈ユーザ情報レコード〉::=〈レコード固有情報〉¥t〈レコード本体〉¥n

〈レコード固有情報〉::=〈レコードID〉'|'〈レコード識別〉'|'〈参照IDリスト〉'|'〈覚書〉

〈レコード本体〉::=〈単語レコード〉|〈対訳レコード〉|〈共起レコード〉|
〈概念体系レコード〉|〈概念記述レコード〉|〈概念見出しレコード〉

〈レコードID〉::=〈１０進整数〉

〈レコード識別〉::=〈１０進整数〉
〈参照IDリスト〉::=〈参照ID〉|〈参照ID〉;〈参照IDリスト〉
〈参照ID〉::=〈１０進整数〉
〈覚書〉::=〈文字列〉

なお，レコード固有情報において，情報が無い場合には，長さ0の文字列を格納する．ただし，これが許されるのは，〈参照IDリスト〉および〈覚書〉のみである．

7.2.2 レコード本体の構造

レコード本体では，基本的に辞書レコード中の情報をそのまま格納する．ただし，品詞，左右連接情報，概念関係子については，辞書レコード中の数値コードに対応するコード文字列を格納する．また，各辞書レコード本体においては，概念識別子のみを格納するかわりに，以下の形式による意味情報を格納する．

〈意味情報〉::=〈概念見出しデータ〉|?〈概念識別子〉
〈概念見出しデータ〉::=〈概念識別子〉¥t〈日本語概念見出し〉¥t〈英語概念見出し〉¥t
〈日本語概念説明〉¥t〈英語概念説明〉
〈概念識別子〉::=〈１６進整数〉
〈日本語概念見出し〉::=〈文字列〉
〈英語概念見出し〉::=〈文字列〉
〈日本語概念説明〉::=〈文字列〉
〈英語概念説明〉::=〈文字列〉

通常は，意味情報として概念見出しデータを格納するが，概念見出し辞書にエントリのない概念識別子の場合には，"?〈概念識別子〉" の文字列として格納する．
なお，レコードの各項目において，情報が無い場合には長さ0の文字列を格納する．

単語レコード

〈単語レコード〉::=〈単語見出し〉¥t〈読み〉¥t〈発音情報〉¥t〈品詞〉¥t
〈左連接属性〉¥t〈右連接属性〉¥t〈意味情報〉¥t〈言語識別〉¥t
〈慣用句フラグ〉¥t〈レコード種別〉¥t〈文法情報〉¥t〈構文木〉
〈単語見出し〉::=〈文字列〉
〈読み〉::=〈文字列〉
〈発音情報〉::=〈文字列〉
〈品詞〉::=〈文字列〉

〈左連接属性〉 ::= 〈文字列〉
〈右連接属性〉 ::= 〈文字列〉
〈言語識別〉 ::= J | E
〈慣用句フラグ〉 ::= 1 | 0
〈レコード種別〉 ::= 1 | 0
〈文法情報〉 ::= 〈文字列〉
〈構文木〉 ::= 〈文字列〉

対訳レコード

〈対訳レコード〉 ::= 〈単語見出し〉¥t〈読み〉¥t〈品詞〉¥t〈言語対識別〉¥t
〈レコード種別〉¥t〈意味情報〉¥t〈対訳情報〉

〈単語見出し〉 ::= 〈文字列〉
〈読み〉 ::= 〈文字列〉
〈品詞〉 ::= 〈文字列〉
〈言語対識別〉 ::= "JE" | "EJ"
〈レコード種別〉 ::= 1 | 0
〈対訳情報〉 ::= 〈文字列〉

共起レコード

〈共起レコード〉 ::= 〈受け側単語見出し〉¥t〈受け側表記情報〉¥t〈受け側品詞〉¥t
〈共起関係子〉¥t〈関係単語表記〉¥t
〈係り側単語見出し〉¥t〈係り側表記情報〉¥t〈係り側品詞〉¥t
〈受け側意味情報〉¥t〈受け側慣用句フラグ〉¥t〈概念関係子〉¥t
〈関係子方向〉〈係り側意味情報〉¥t〈係り側慣用句フラグ〉¥t
〈言語識別〉¥t〈表層共起頻度〉¥t〈共起項目頻度〉¥t
〈受け側共起要素頻度〉¥t〈係り側共起要素頻度〉¥t〈例文〉

〈受け側単語見出し〉 ::= 〈文字列〉
〈受け側表記情報〉 ::= 〈文字列〉
〈受け側単語品詞〉 ::= 〈文字列〉
〈共起関係子〉 ::= 〈文字列〉
〈関係単語表記〉 ::= 〈文字列〉
〈係り側単語見出し〉 ::= 〈文字列〉
〈係り側表記情報〉 ::= 〈文字列〉
〈係り側単語品詞〉 ::= 〈文字列〉
〈受け側意味情報〉 ::= 〈意味情報〉

〈受け側慣用句フラグ〉 ::= 1 | 0
〈概念関係子〉 ::= 〈文字列〉
〈関係子方向〉 ::= 1 | 0
〈係り側意味情報〉 ::= 〈意味情報〉
〈係り側慣用句フラグ〉 ::= 1 | 0
〈言語識別〉 ::= J | E
〈表層共起頻度〉 ::= 〈１０進整数〉
〈共起項目頻度〉 ::= 〈１０進整数〉
〈受け側共起要素頻度〉 ::= 〈１０進整数〉
〈係り側共起要素頻度〉 ::= 〈１０進整数〉
〈例文〉 ::= 〈文字列〉

概念体系レコード

〈概念体系レコード〉 ::= 〈下位概念〉¥t〈上位概念〉

〈下位概念〉 ::= 〈意味情報〉
〈上位概念〉 ::= 〈意味情報〉

概念記述レコード

〈概念記述レコード〉 ::=
　　〈始点概念〉¥t〈終点概念〉¥t〈概念関係子〉¥t〈真偽値〉¥t〈記述区分〉

〈始点概念〉 ::= 〈意味情報〉
〈終点概念〉 ::= 〈意味情報〉
〈関係子〉 ::= 〈文字列〉
〈真偽値〉 ::= 1 | 0
〈記述区分〉 ::= 1 | 0

概念見出しレコード

〈概念見出しレコード〉 ::= 〈意味情報〉