6S−6

大規模電子化辞書開発における高機能辞書エディタ
−　そのデータ構造　−

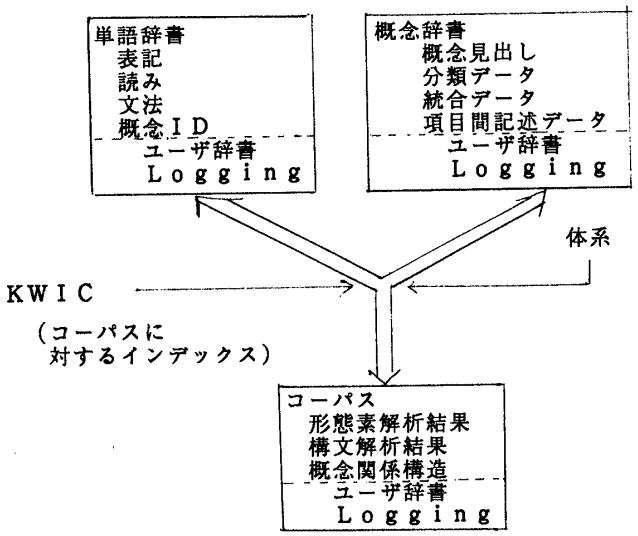
三輪和弘、中沢正幸、鈴木美穂、安達久博*
株式会社　日本電子化辞書研究所
株式会社　東芝

はじめに

（株）日本電子化辞書研究所（ＥＤＲ）では、実際の文章データに基づいて辞書開発を行っている。つまり、辞書のすべての情報は少なくとも一つ以上の文章によって保証された情報である。この保証を常に維持することは、辞書情報あるいは文章データの変更を他のデータへ反映しなければならないことを意味する。ＥＤＲ辞書は現段階においても約４００メガバイトを有し、コーパスデータ及びＫＷＩＣデータは約３０ギガバイト近くになる予定である。
したがってこの保証を実現するためにＥＤＲでは、高機能辞書エディターを開発している。高機能辞書エディターは、上記データをスムーズに参照しエディティングするために開発されている。
また、設計思想としては、Ｈｙｐｅｒ　Ｔｅｘｔ及びオブジェクト思考を導入している。
さらに、概念辞書の開発において、概念間のリンクを行う概念統合、概念の抽象化を行う概念体系の開発にも高機能辞書エディターが不可欠である。
本論文では、高機能辞書エディターの対象となるそれぞれのデータの構造とデータ間の関係、及びそのアクセス方法に付いて報告する。

１．　ＥＤＲ辞書データとコーパスデータの関係について

ＥＤＲ辞書とコーパスは、単語や概念ＩＤによってリンクが張られている。したがって、辞書データあるいはコーパスデータの修正はそのデータのみでなく、他のデータをも修正したことになる。また，それぞれのデータの内部においても相互に依存しあっているので１カ所の修正は他と矛盾を起こすことがある。これらの管理も行わなければならない。データ間の関係は、図−１のデータを辿ることにより参照される。それぞれのデータは，基本的な辞書情報としてのデータと高機能辞書エディター内の情報としての修正履歴及び他のユーザとの整合性を取るためのユーザ辞書（一時的な修正部分の格納場所）とから構成されている。



図−１　高機能辞書エディターが対象とするデータの構成図

２．　ＥＤＲ辞書データとコーパスデータのデータ構造について

ＥＤＲ辞書データのデータ構造
ＥＤＲ辞書は、単語辞書と概念辞書の２つから構成されている。単語辞書は形態そ解析、構文解析に利用できる文法情報が格納されている。また単語辞書のつのエントリーは、１つの概念に対応している。その概念にはユニークなＩＤと他の概念と区別できる程度の文章表現が書かれておりそれを概念見出しと呼んでいる。ただし、概念見出しは単語辞書ではなく概念辞書に記述されている。概念辞書は、概念見出しの他に体系の実データとなる分類データ、概念の発散を抑えるための統合データ、概念間の関係を記述した関係データ、概念項目間の関係を記述した項目間記述データといった概念に関係したデータが概念ＩＤによって格納されている。
それぞれのデータに対応するクラスとその参照されるデータを以下に示す。

High Quality Editor for Developing Electronic Dictionary
Hisahiro ADACHI, Kazuhiro MIWA, Hasayuki NAKAZAWA and Miho SUZUKI
Japan Electronic Dictionary Research Institute, Ltd.

単語辞書エントリに対応するクラスＷｏｒｄＣｌａｓｓ、概念辞書エントリに対応するクラスＣｏｎｃｅｐｔＣｌａｓｓ、コーパスデータのクラスとそのデータ構成以下のようになっている。

```
ＷｏｒｄＣｌａｓｓ
        単語表記        :  単語の表記
        読み            :  単語の読み
        連接属性        :  形態素解析で利用するための文法属性
        文法情報        :  構文解析で利用される文法属性
        概念ＩＤ        :  固有の概念に付けられるＩＤ
ＣｏｎｃｅｐｔＭｉｄａＣｌａｓｓ
        概念ＩＤ        :  概念ＩＤ
        概念見出し      :  他の概念と明確に区別できうる程度の自然言語表記
ＣｏｎｃｅｐｔＲｅｌＣｌａｓｓ
        概念ＩＤ１      :  概念見出しに対応したＩＤ
        関係            :  概念ＩＤ１と概念ＩＤ２の間に成り立つ関係
        概念ＩＤ２      :  他の概念見出しに対応したＩＤ
ＣｏｎｃｅｐｔＣＰＳ１ｓｔＣｌａｓｓ
        形態素解析結果  :  概念ＩＤリスト
ＣｏｎｃｅｐｔＣＰＳ２ｎｄＣｌａｓｓ
        構文解析結果    :  係り受け構造データ
ＣｏｎｃｅｐｔＣＰＳ３ｒｄＣｌａｓｓ
        概念関係構造    :  概念関係構造
```

対訳リンクや分類データ及び統合データ，概念記述データはすべてこの関係として表現されている。

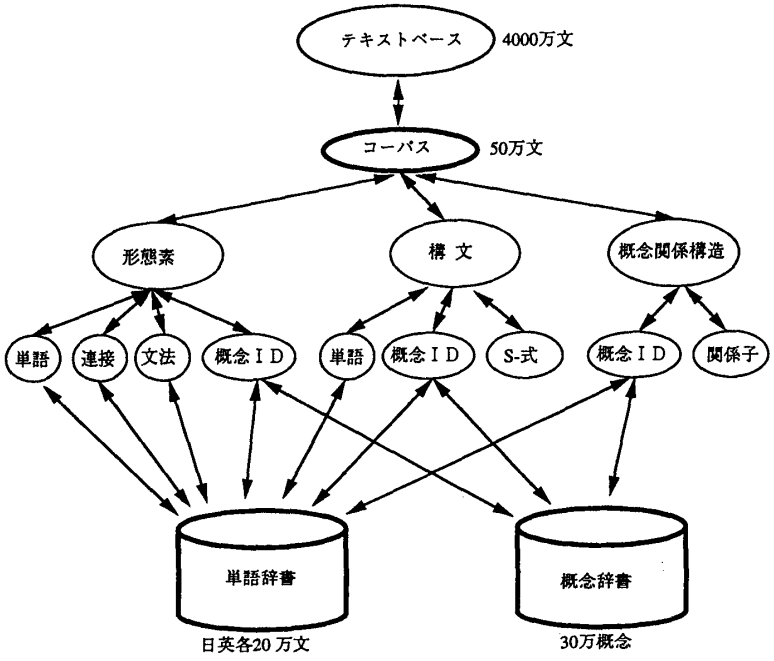
コーパスデータは、ＥＤＲ辞書により解析され人手で検証され、その文分析結果として得られた形態素解析結果と構文解析結果、概念関係構造から構成されている。内部データとしては、単語表記、概念ＩＤのリスト及び関係構造を持つクラスのデータとして登録される。

３．　ＥＤＲ辞書データとコーパスデータのアクセスについて

図－１で示したようにそれぞれのデータは、自分自身の情報とリンク情報を持っている。ユーザは常にすべてのデータを参照したいわけではない。それどころか参照したいのは部分情報の方が多い。

そこで我々は、データの持つ情報を属性とし、エントリーごとのデータの取出しをクラスメソッドとして定義した。これによってユーザは、自分の欲しい情報を持つインスタンスを生成しデータにアクセスすることができる。それぞれのクラス内の情報の取出しは，それぞれのクラスへのメッセージによって行われる。

以下にその関係を示す。



ある概念の分類項目が知りたい場合は以下のようなクラスの組み合わせになる。

```
WordClass(単語, ConceptID);
ConceptRelClass(3,ConceptID,Label);
```

まとめ

種々の観点から辞書及び参照文章をエディティングするためのデータの構造とアクセス方法を実現できた。今後は１つのデータの変更が他にどのような影響を及ぼすかユーザが把握できるシミュレーション機能を検討中である。

謝辞

最後に、この研究の機会与えて下さったＥＤＲの横井所長、第四研究室吉田室長、及び終始貴重なご意見を頂いた第一研究室内田室長並びに研究員の皆様に感謝致します。

参考文献：

［１］　ＥＤＲ電子化辞書: EDR TR-016 November 1989
［２］　辞書開発支援システム: EDR TR-015 November 1989
［３］　Joan Peckham,Fred Maryanski: "Semantic Data Models" ACM Computing Surveys Vol.20, No3