「現代日本語書き言葉均衡コーパス」のための形態論情報データベースについて

On the Morphological Information Database for Balanced Corpus of Contemporary Written Japanese

小木曽 智信　　中村 壮範

Toshinobu Ogiso　　Takenori Nakamura

国立国語研究所 言語資源研究系，立川市緑町 10-2

National Institute for Japanese Language and Linguistics, Department of Corpus Studies,

10-2 Midori-cho, Tachikawa city, Tokyo

あらまし：国立国語研究所を中心に構築が進められている「現代日本語書き言葉均衡コーパス」のために開発した形態論情報データベースについて報告する。このデータベースは辞書データベースとコーパスデータベースとからなり、語彙表を介してお互いのデータの同期がとられている。辞書データベースの内容は形態素解析辞書 UniDic のソースデータとして利用される。コーパスデータベースは形態論情報を埋め込んだ XML 文書を出力する。

Summary: This paper reports the morphological information database for the Balanced Corpus of Contemporary Written Japanese, which is under way at the National Institute for Japanese Language and Linguistics. This database consists of dictionary database and corpus database, and two databases are synchronized via the lexicon table. Contents of dictionary database are used for the source data of UniDic, an electrical dictionary for morphological analysis. Corpus database can output XML documents annotated with morphological information.

キーワード：コーパス，形態素解析，辞書，データベース

Keywords: corpus, morphological analysis, dictionary, database

1. はじめに

本発表では、『現代日本語書き言葉均衡コーパス』（以下 BCCWJ）のために開発した「形態論情報データベース」の設計と実装について報告する。BCCWJ は総語数 1 億を目標に、人間文化研究機構国立国語研究所コーパス開発センターを中心に構築が進められている大規模なコーパスであり、2011 年の公開を目指している。このコーパスにはすべてのテキストに長短二種類の形態論情報が付与され、語を基本とした検索や集計が可能となる。

形態論情報データベースは、このコーパスの形態論情報を付与し整備するとともに、コーパスを利用した研究を可能にするためのデータベースである。国立国語研究所で開発・運用を行っている。

このデータベースは辞書データベースとコーパスデータベースからなり、語彙表を介してお互いのデータの同期がとられている。辞書データベースは、形態素解析辞書 UniDic の元となる見出し語のデータを格納しており、コーパス中に現れる新語を随時追加している。コーパスデータベースは UniDic によって BCCWJ のテキストを解析した結果を取り込んだものであり、形態論情報の人手による修正を行うとともに、様々な手段によって検索することが可能となっている。

形態論情報を格納したコーパスのデータベースは、全文テキスト検索が可能であると同時に、当該のテキストの構成要素であるすべての語について詳細な属性（形態論情報）を保持し、これらを組み合わせた検索にも対応する必要がある。また、データの整合性を保つために、辞書データベースと関連づけて管理を行う必要がある。

こうした点で、一般的なデータベースとは異なる特徴を持っている。

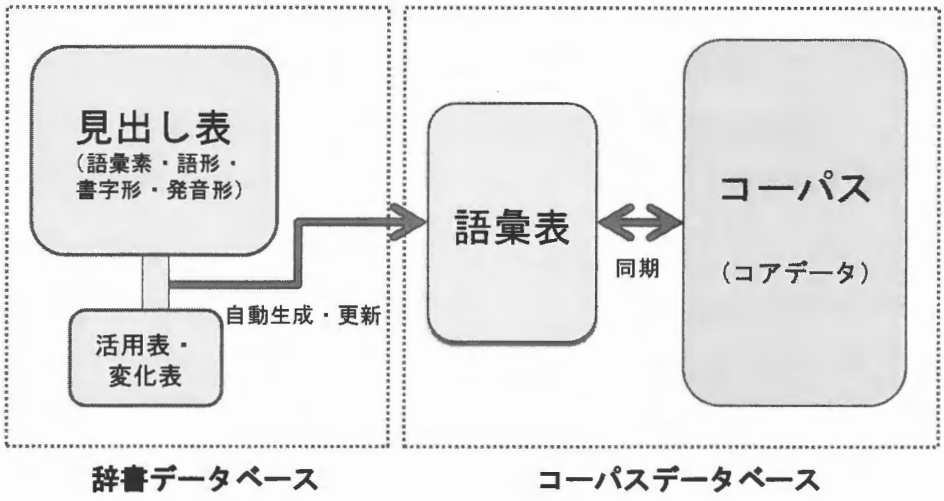

図1　形態論情報データベース全体図

2. 形態論情報データベースの概要

形態論情報データベースの主な利用目的は、次の3点である。

(1) 形態素解析辞書 UniDic の元となる見出し表・活用表を格納し、見出し語の追加・修正作業を行う

(2) BCCWJ の短単位で解析されたテキストを格納し、人手による修正を行って精度を高めたデータ(コアデータ)を作成する

(3) 短単位で解析されたテキストを格納し、コーパスを利用した研究に利用する

(1)は辞書の見出し語、(2), (3)はコーパスのデータを扱うことになる。これに対応して、形態論情報データベースは、(1)の辞書見出しを格納する「辞書データベース」と(2), (3)のコーパスを格納する「コーパスデータベース」に分かれている。コーパスの形態論情報と辞書の情報を同一に保つ必要があるため、それぞれのデータベースは中間に辞書見出し表から生成される「語彙表」を挟んで連係している。コーパスに出現したすべての語は、原則として語彙表のいずれかのレコードと関連付けられる。(図 1)

形態素解析辞書の作成という観点から見たときには、(1), (2)は形態素解析辞書 UniDic の元となるデータを用意するための作業に相当する。(1)の見出し表を組み合わせることにより解析辞書の見出し表(辞書)が生成され、(2)のコアデータから学習用コーパスが作られる。この二つのデータ元に、機械学習により形態素解析辞書が作成される。

(3)はこの形態素解析辞書によって解析されたテキストデータを学習コーパスと同様の形式で格納したものである。このデータは言語研究に利用するだけでなく、辞書の整備(未登録の語を見つけ出し追加する等)のためにも利用される。

「形態論情報データベース」は、データベースソフト(DBMS)に Microsoft SQL Server を、クライアントに Microsoft Access で作成した専用アプリケーションを用いるクライアント・サーバ型のシステムとして構築されている。サーバの OS には Windows 2003 Server R2 Standard、データベース管理システム(DBMS)として Microsoft SQL Server 2005 Standard Edition (SP2) を利用している。十分なメモリを利用するためどちらも 64 ビット版 (x64 Edition) を利用している。

3. 辞書データベース

辞書データベースは、形態素解析辞書 UniDic の元となる見出し語のデータベースである。見出し語のテーブルのほか、活用表などの辞書作成に必要な情報からなる。

辞書データベースの基本となる見出し表は、UniDic の見出し設計にあわせて作成された「短単位語彙素」、「短単位語形」「短単位書字形」「短単位発音形」の 4 つである。

UniDic では図 2のような階層化された見出し語が設定されている(伝ほか 2007)。

「語彙素」は国語辞典の見出し語に相当するレベルで、語の意味や語の出自などの情報はここに記述される。「語形」は異語形を区別するレベルで、たとえば「アマリ(余り)」に対する「アンマリ」「アンマシ」「アンマ」といった異語形、上一段活用と文語上二段活用といった活用の違いのほか、可能動詞形もここで区別される。「書字形」は異表記を区別するレベルで、漢字を使うか仮名書きするかといった違いのほか、送り仮名の揺れ

もここに記述される。「発音形」には発音やアクセントなどの情報が記述される。

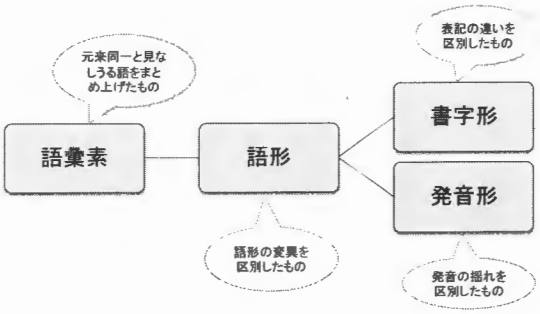

図2　UniDic の見出し設計

辞書データベースの見出し表はこの階層をそのまま反映し、「短単位語彙素テーブル」「短単位語形テーブル」「短単位書字形テーブル」「短単位発音形テーブル」の4つからなっている(図 4)。

各見出し語は、具体的には図 3のように階層化された形で格納されることになる(発音形は語形から直接結合する)。

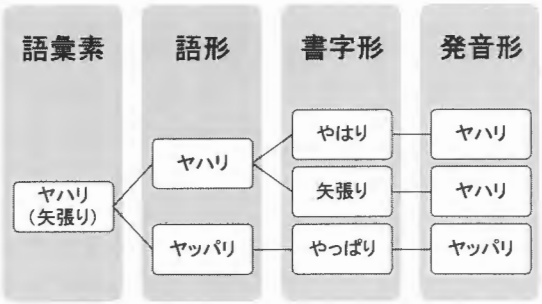

図3　UniDic の見出し構造の例

見出し表

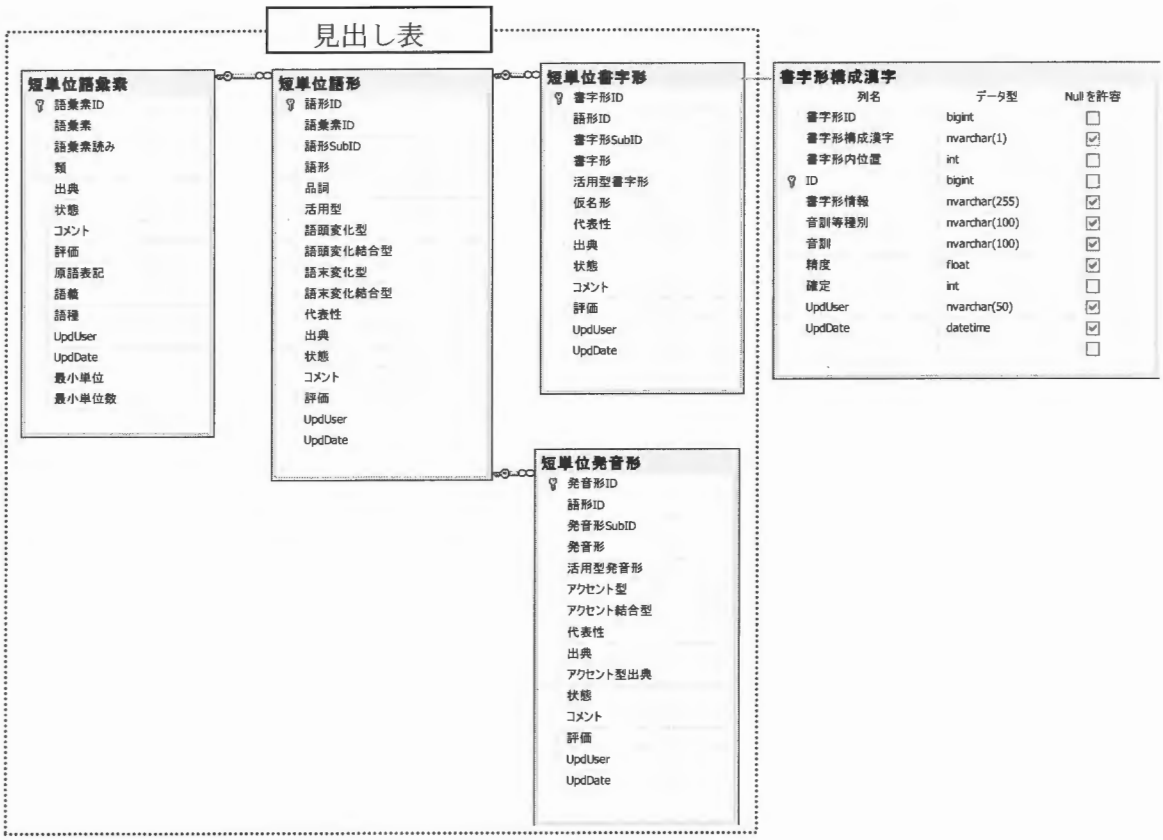

図4　辞書データベースの見出し表

語彙表の展開

辞書データベースには、見出し表のほかに、活用語を展開するための「活用表」と「活用型表」「活用形表」、語頭変化形を展開するための「語頭変化表」、語末変化形を展開するための「語末変化表」が存在する。

語形は、語頭変化(連濁など)・語末変化(促音化など)・活用の3種の変化により語形を派生する。個々の語形は、語頭変化・語末変化・活用の順に反映され、出現形に展開されることになる(図 5, 図 6)。

図5　語彙表の展開（例：辛い）

結果　メッセージ

	lid	lemma	reading	type	meaning	form	pos	cType	sysCType	cForm	origin	pronBase	orthBase	kanaBase	pronToken	orthToken	kanaTok...	aModType
1	10853287968514785	読む	ヨム	用		ヨム	動詞一般	五段-マ行一般	五段-マ行	仮定形一般	lcpr	ヨム	読む	ヨム	ヨメ	読め	ヨメ	
2	10853287968514786	読む	ヨム	用		ヨム	動詞一般	五段-マ行一般	五段-マ行	仮定形-融合	lcpr	ヨム	読む	ヨム	ヨミャ	読みゃ	ヨミャ	
3	10853287968514849	読む	ヨム	用		ヨム	動詞一般	五段-マ行一般	五段-マ行	命令形	lcpr	ヨム	読む	ヨム	ヨメ	読め	ヨメ	
4	10853287968514659	読む	ヨム	用		ヨム	動詞一般	五段-マ行一般	五段-マ行	意志推量形	lcpr	ヨム	読む	ヨム	ヨモ	読も	ヨモ	M1@0
5	10853287968514658	読む	ヨム	用		ヨム	動詞一般	五段-マ行一般	五段-マ行	意志推量形	lcpr	ヨム	読む	ヨム	ヨモッ	読もっ	ヨモッ	M1@1
6	10853287968514657	読む	ヨム	用		ヨム	動詞一般	五段-マ行一般	五段-マ行	意志推量形	lcpr	ヨム	読む	ヨム	ヨモー	読もう	ヨモウ	M1@1
7	10853287968514625	読む	ヨム	用		ヨム	動詞一般	五段-マ行一般	五段-マ行	未然形一般	lcpr	ヨム	読む	ヨム	ヨマ	読ま	ヨマ	
8	10853287968514731	読む	ヨム	用		ヨム	動詞一般	五段-マ行一般	五段-マ行	終止形一般	lcpr	ヨム	読む	ヨム	ヨム	読む	ヨム	NULL
9	10853287968514753	読む	ヨム	用		ヨム	動詞一般	五段-マ行一般	五段-マ行	連体形一般	lcpr	ヨム	読む	ヨム	ヨム	読む	ヨム	NULL
10	10853287968514689	読む	ヨム	用		ヨム	動詞一般	五段-マ行一般	五段-マ行	連用形一般	lcpr	ヨム	読む	ヨム	ヨミ	読み	ヨミ	
11	10853287968514693	読む	ヨム	用		ヨム	動詞一般	五段-マ行一般	五段-マ行	連用形-撥音便	lcpr	ヨム	読む	ヨム	ヨン	読ん	ヨン	
12	10853305148384033	読む	ヨム	用		ヨム	動詞一般	文語四段-マ行	文語四段-マ行	命令形	活	ヨム	読む	ヨム	ヨメ	読め	ヨメ	
13	10853305148384001	読む	ヨム	用		ヨム	動詞一般	文語四段-マ行	文語四段-マ行	已然形	活	ヨム	読む	ヨム	ヨメ	読め	ヨメ	
14	10853305148383841	読む	ヨム	用		ヨム	動詞一般	文語四段-マ行	文語四段-マ行	意志推量形	活	ヨム	読む	ヨム	ヨモー	読まう	ヨマウ	M1@1
15	10853305148384060	読む	ヨム	用		ヨム	動詞一般	文語四段-マ行	文語四段-マ行	未然形+省略	活	ヨム	読む	ヨム	ヨマ	読	ヨマ	NULL
16	10853305148383809	読む	ヨム	用		ヨム	動詞一般	文語四段-マ行	文語四段-マ行	未然形一般	活	ヨム	読む	ヨム	ヨマ	読ま	ヨマ	
17	10853305148383915	読む	ヨム	用		ヨム	動詞一般	文語四段-マ行	文語四段-マ行	終止形一般	活	ヨム	読む	ヨム	ヨム	読む	ヨム	NULL
18	10853305148383937	読む	ヨム	用		ヨム	動詞一般	文語四段-マ行	文語四段-マ行	連体形一般	活	ヨム	読む	ヨム	ヨム	読む	ヨム	NULL
19	10853305148384039	読む	ヨム	用		ヨム	動詞一般	文語四段-マ行	文語四段-マ行	連用形+省略	活	ヨム	読む	ヨム	ヨミ	読	ヨミ	NULL
20	10853305148383875	読む	ヨム	用		ヨム	動詞一般	文語四段-マ行	文語四段-マ行	連用形-ウ音便	活	ヨム	読む	ヨム	ヨー	読う	ヨウ	
21	10853305148383873	読む	ヨム	用		ヨム	動詞一般	文語四段-マ行	文語四段-マ行	連用形一般	活	ヨム	読む	ヨム	ヨミ	読み	ヨミ	
22	10853305148384043	読む	ヨム	用		ヨム	動詞一般	文語四段-マ行	文語四段-マ行	連用形-撥音便	活	ヨム	読む	ヨム	ヨン	読	ヨン	NULL
23	10853305148383877	読む	ヨム	用		ヨム	動詞一般	文語四段-マ行	文語四段-マ行	連用形-撥音便	活	ヨム	読む	ヨム	ヨン	読ん	ヨン	

図6　展開された語彙表の例

4. コーパスデータベース

BCCWJ のデータは XML で記述されている。コーパスデータベースでは、この情報を関係データベースの一般的な表で表現するために、Stand-off Annotation の方法により、「文字表」「短単位表」「文字修正表」「数字タグ表」「ルビ表」「タグ表」の各表に分けて取り込んでいる。形態論情報の処理に直接関連するタグのみ専用テーブルに書き込み、その他のタグは一括してタグ表で保管する。いずれのテーブルもサンプル ID と原文における文字位置をキーとして関連付けられている。

コーパスデータベースには各種のコーパスが格納されている。そのうち、人手修正を施したデータをコアデータと呼ぶ。コアデータは形態素解析辞書 UniDic の学習用コーパスとして利用される。コアデータ以外のデータは、見出し表に登録するための未登録語の採集や、コーパスを利用する研究のために用いるデータである。

コーパスデータベース内のテーブルは文字テーブルを軸として、サンプル ID と文字開始位置・文字終了位置をキーにして関連付けされている。また、辞書データベースとは語彙表テーブルを介して関連付けされている。これによりコーパスデータベース用のアプリケーションからはコーパスデータベース内の全てのデータにアクセスできるようになっている。

短単位テーブル

短単位テーブルは形態素解析結果を取り込んだもので、コーパスデータベース内でも最も重要な役割をもつテーブルである。UniDic のもつ豊富な形態論情報のほか、コーパス中の位置情報をもつ。形態論情報は、「語彙素」「語彙素読み」「品詞」「活用型」「活用形」「書字形」「発音形」でユニークとなり語彙表中の1エントリと対応し、辞書データベースと関連づけられる。位置情報は、サンプル ID とサンプル中の連番で一意に特定される。

長単位テーブルと文節テーブル

長単位は、BCCWJ の形態論情報として付与される言語単位の一つで、文節をもとに、そこから付属語等

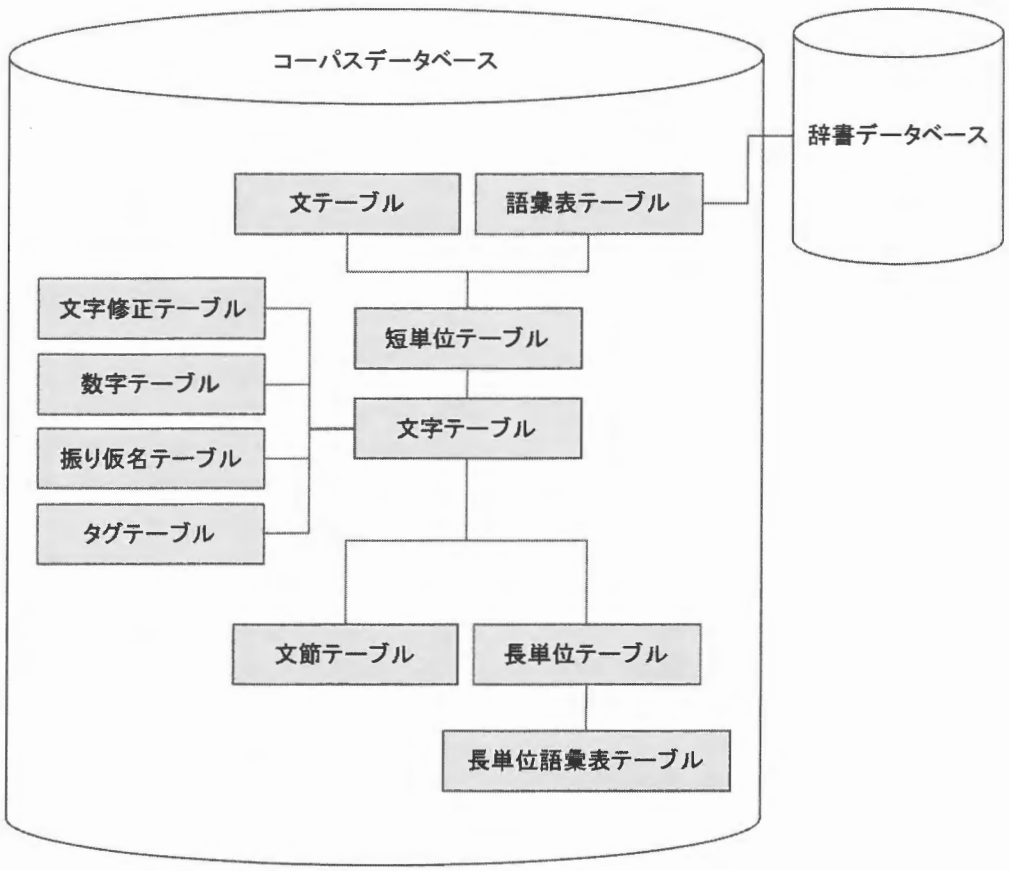

図7　コーパスデータベースのテーブル

を切り出したものに相当する。一つの長単位は、一つの短単位または複数個の短単位の連続となる。

短単位と長単位・文節は、表 1のような関係にある(表中のBは境界位置であることを示す)。文節境界は常に長単位境界であり、文節・長単位境界は常に短単位境界となる。また、文節や長単位は短単位の連続からなる。ただし、注釈的な括弧などにより、長単位が短単位の連続とならない場合がある。短単位と長単位は、語彙素・品詞・活用型等の情報をもつが、文節は境界のみを記録している。

表1　短単位・文節境界・長単位の例

短単位境界	短単位	文節境界	長単位境界	長単位
B	文化	B	B	文化庁文化交流使事業
B	庁			
B	文化			
B	交流			
B	使			
B	事業			
B	は		B	は
B	,		B	,
B	芸術	B	B	芸術家
B	家			
B	,		B	,
B	文化	B	B	文化人等
B	人			
B	等			
B	,		B	,
B	文化	B	B	文化
B	に			に
B	携わる	B	B	携わる
B	人々	B	B	人々
B	に			に
B	,		B	,
B	一定	B	B	一定期間
B	期間			

長単位はコーパスに出現したものを単位として認めるという形を取っており、コーパスから切り離した見出し表としては管理しない。そのため形態論情報データベースではコーパスデータベースの中でのみ取り扱われ、辞書データベースとは直接関係しない。

長単位に関係するテーブルとしては、長単位テーブルと文節テーブル、長単位語彙表テーブルがある。
長単位テーブルは、出現した長単位の情報を格納するテーブルであり、語彙素・品詞・活用型などの情報が、短単位の情報を利用して付与される。文節テーブルは文節境界を記録するためのテーブルである。長単位テーブルと文節テーブルは、短単位テーブルと同様に、文字表テーブルのサンプル ID と文字開始位置・文字終了位置によって関連付けされている。

5. クライアントアプリケーション

辞書データベースとコーパスデータベースは、それぞれ Microsoft Access による専用のクライアントアプリケーションによって修正・管理される。辞書管理ツール「UniDic Explorer」とコーパス修正ツール「大納言」である。

辞書管理ツール・UniDic Explorer

「UniDic Explorer」は辞書データベースに見出し語を追加するための中心となるツールである。見出し語の追加・修正作業には、見出し語表の階層をそのまま表示し、修正することが可能となっている。また、階層構造を保ったまま、見出し語をコピー・移動することができる。各見出し語はコーパス中の用例と関連づけられており、コーパス中の頻度が表示されるほか、用例の参照も可能である(図 8)。

コーパス修正ツール・大納言

「大納言」はコーパスの検索・修正を行うためのツールである。

コーパスの検索は、「短単位」を用いて、その語彙素読み・語彙素・書字形などの形態論情報を用いた検索ができるほか、全文テキスト検索も可能となっている。全文検索は、短単位テーブルから生成した文テーブルにインデックスを付与することにより実現している。さらに、前後に出現する短単位の形態論情報を複数組み合わせた「高度な検索」機能をもつ。

コーパスの修正は、短単位表については、語彙表を参照しながら、辞書データベースの見出し語と関連づけながら修正を行うシステムとなっている。これにより、常にコーパスと辞書との整合性を保ちつつ、随時修正を行うことができるようになっている。

短単位のほかに、長単位や文節境界の修正、誤った本文テキストの修正など、コーパスの各種テーブルについて、関連する単位との整合性を保ちつつ修正を行うことが可能となっている。

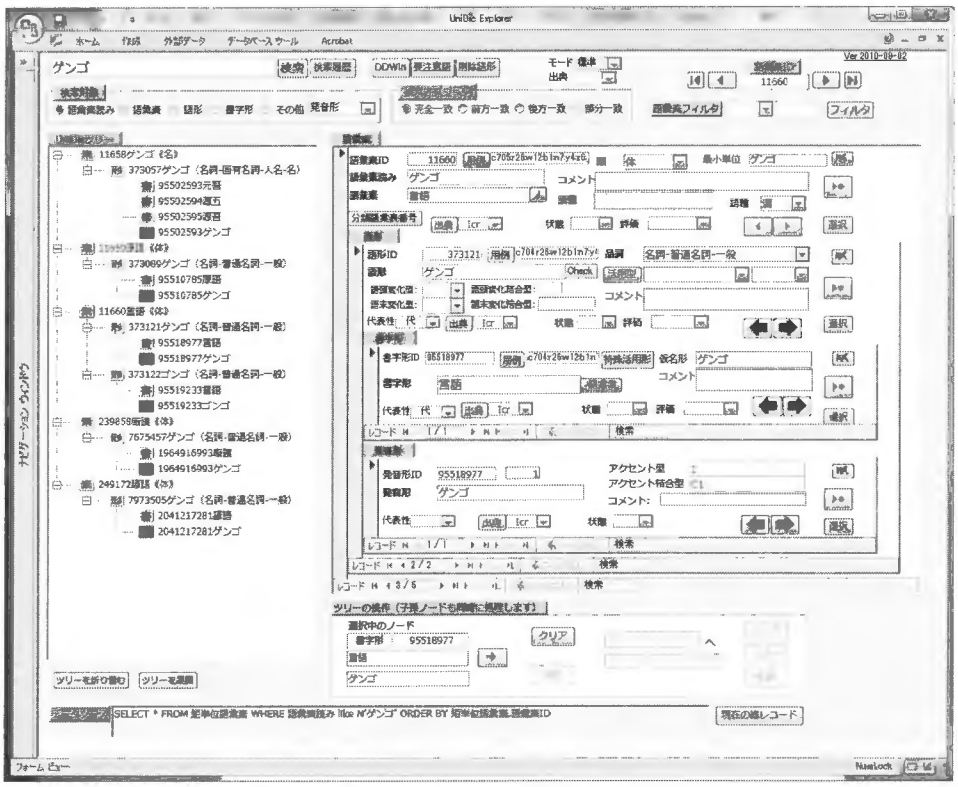

図8　辞書管理ツール・UniDic Explorer

図 9　コーパス修正ツール・大納言

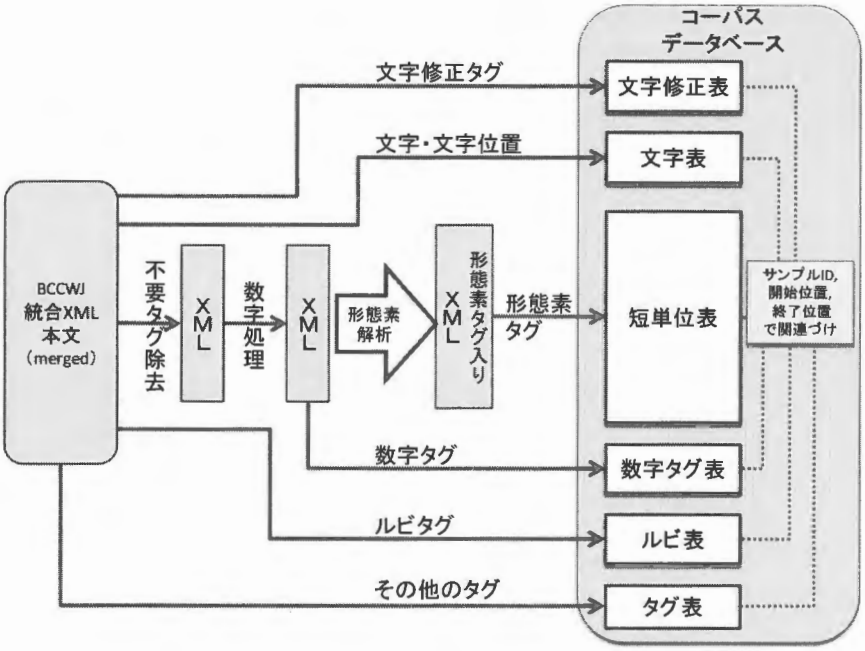

図10　BCCWJ サンプルの形態素解析とインポート

6. データのインポートとエクスポート

BCCWJ のサンプルに形態素解析を施し、コーパステーブルにインポートする際には、図 10に示すような流れになる。形態素解析や数字処理の邪魔になるタグの除去や、数字変換などの処理が加わるため、それぞれの表の情報を取り出す段階が異なる。タグ・文字テーブルは、元の XML 文書から直接取り出す(したがって、文字テーブルとタグテーブルからXML 文書が完全に再現できる)。数字タグは数字処理後のデータから取り出すことになる。

取り込んだデータは、人手で修正した後、元の XML 文書に形態素タグを埋め込んだ XML 形式でエクスポートすることができる。DBMS の管理ツール(SQL Server Management Studio)上で、SQL 文を実行することよって出力される。エクスポート用の SQL 文では、各テーブルを結合し、データベース内部で XML 型のデータとして生成した後、ファイル出力している。データベース内でXML型のデータを生成するため、この時点で整形式の XML であることが保証される。

テーブルの結合時には、タグテーブルを参照するが、このとき、ルビや数字などの別テーブルで管理されているタグはタグテーブルから出力せず、各テーブルの情報を元にタグを再構成して出力することになる。

7. おわりに

BCCWJ のために開発した「形態論情報データベース」の概要を紹介した。大規模コーパスの構築には、注意深く設計されたデータベースが必須である。また、豊富な情報を持ち、日々人手修正が加わる1億語規模のコーパスデータベースを運用していくことは容易ではなく、多くの経験を積むこととなった。こうしたノウハウを、今後のより大規模なコーパスや、歴史コーパスなどの構築に活かしていきたいと考えている。

参考文献

Masaaki Asahara, Ryuichi Yoneda, Akiko Yamashita, Yasuharu Den and Yuji Matsumoto (2002) Use of XML and relational databases for consistent development and maintenance of lexicons and annotated corpora , In *Proceedings of the Third International Conference on Language Resource and Evaluation* (LREC 2002), Las Palmas, Canary Islands, Spain, pp.1372-1378.

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵(2007)「コーパス日本語学のための言語資源:形態素解析用電子化辞書の開発とその応用」(『日本語科学』22 号, pp.101-122)

小木曽智信・中村壮範(2009)『『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装』国立国語研究所内部報告書(LR-CCG-08-04)

参考 URL

形態素解析辞書 UniDic http://download.unidic.org