# 汎電子化辞書：表層レベルの構造

横井 俊夫　木村 和広　小泉 敦子　三吉 秀夫　川田 亮一
日本電子化辞書研究所

## 1　はじめに

基本構造の表層の記述レベルにそって汎電子化辞書 [1] の情報構造を述べる。語表層辞書、文表層辞書（とりあえず句も含める）、文章表層辞書、文書表層辞書の4サブ辞書の内部構造と相互の関係である。言語の種類については日本語と英語を中心にする。

情報構造の記述にあたっては、特定の検索機構、処理機構、推論機構を前提としない。あくまでも、情報の論理的関係の記述である。便法として、簡略化された拡張 BNF と自然言語を用いて説明する。

(1) ⟨AAA⟩ :BBB　AAA という名前の項目を定義。BBB は自然言語文章等による項目の説明、あるいはコメント。
(2) ⟨CCC⟩...　CCC という項目がひとつ以上複数繰り返される。
(3) ⟨DDD⟩　項目 DDD がサブ項目（下位要素）EEE と FFF を持つ。
⟨EEE⟩
⟨FFF⟩

## 2　表層辞書の構造

### 2.1　基本構造

表層辞書における情報の基本単位は外見的に他と明確に区別しうる表記を持ち、上位の構造の中でどのような要素に振舞うのか明確に定義でき、指示している内容（意味、対象）を確定でき、使用される環境（状況）を規定できるという4つの性質を持つものを基本単位とする。

| | |
|---|---|
| ⟨表層辞書⟩ | |
| ⟨表層辞書項目⟩... | :⟨表記情報⟩によって順序付けがなされる |
| ⟨表記情報⟩ | :表記の定義 |
| ⟨構造情報⟩ | |
| ⟨要素カテゴリ⟩ | :上位の構成構造においてどのような要素となるのか |
| ⟨構成構造⟩ | :下位の要素によってどのような構成構造となるのか |
| ⟨指示情報⟩ | :表記が指示する内容 |
| ⟨環境情報⟩ | :運用情報 |

### 2.2　語表層辞書

語に対する上位の構成構造は文の形態素列構造と構文構造（句構造、依存構造等）である。したがって、語の選択にあたっては、まず、単一の形態素であることが優先される。語を構成する形態素には、語の中心的意味を表す語基（いわゆる内容形態素）と機能的・文法的意味を主に表す接辞（いわゆる機能形態素）がある。語基については、語彙体系の中で、中心となるものから周辺に位置するものまで広範囲かつ大量に収録する。接辞はすべて登録すべきである。また、複合語も、それが語として用いられる可能性のあるものは、できるだけ広く収録す

べきである。ただし、複合語については、すべての可能な組合せを網羅することは出来ないので、何らかの基準で登録を制限する必要がある。

| | |
|---|---|
| ⟨語表層辞書⟩ | |
| ⟨語表層辞書項目⟩... | |
| ⟨表記情報⟩ | |
| ⟨代表表記⟩ | :語を代表する表記。その語が実際に使用される表記のバリエーションうち、使用頻度上、あるいは語彙体系上、代表と考えられるもの。 |
| ⟨発音⟩ | :一般にはIPA音標文字で表示。日本語ではカナで代用可。 |
| ⟨異表記リスト⟩ | :表記のバリエーションのリスト。日本語では、漢字の種類、送り仮名の付け方など、英語では、英米の綴り方の違いなど。 |
| ⟨不変化部指定⟩ | :活用語に対し、その不変化部分。国文法の語幹とは必ずしも一致する必要はない。 |
| ⟨語形成⟩ | :語基に対し、どのような派生語、複合語を形成するかを表示。 |
| ⟨構造情報⟩ | |
| ⟨要素カテゴリ⟩ | |
| ⟨形態素情報⟩ | |
| ⟨接続情報⟩ | :語が文中でどのような環境に現れるかを形態素間の接続可否の関係として表示。 |
| ⟨構文情報⟩ | |
| ⟨統語範疇⟩ | :構文的観点から見た時の語のクラス。品詞。通常、品詞分類は形態・構文・意味の各観点を総合して行うことが多いが、ここでは、構文的機能・構文的意味から語および語基を分類したものを品詞分類とする。 |
| ⟨文型情報⟩ | :述語となることのできる語基に対し、それが要求する項の数と格形態を典型的な順序で示す。 |
| ⟨文法範疇⟩ | :文法形式によってまとめられるカテゴリカルな意味と表現の類別を示す。テンス、ヴォイス、アスペクト、ムード。 |
| ⟨構成構造⟩ | |
| ⟨語構造⟩ | :派生語、複合語など複数の形態素からなる語に対し、語の内部構造を表示。 |
| ⟨指示情報⟩ | |
| ⟨語義識別子⟩ | :語義の同一性を示すための識別記号。 |
| ⟨語義⟩ | |
| ⟨語義説明文⟩ | :語義の自然言語による説明文、定義文。 |
| ⟨語概念表記⟩ | :語義を代表する典型的な語。 |
| ⟨文脈⟩ | :用例文の列挙により語義を表現。 |
| ⟨言語内指示⟩ | :同義語、類義語等の列挙により語義を表現。 |
| ⟨言語間指示⟩ | :他言語を意味表現言語とする意味記述。対訳関係。 |
| ⟨言語1対訳リスト⟩ | |
| ⟨言語2対訳リスト⟩ | |
| ⟨環境情報⟩ | :語の使用される分野、頻度、位相、語種等の語の運用に関わる情報。 |

## 2.3 文表層辞書

一般化、体系化されたコーパスである。主要な情報構造は以下のとおり。

| | |
|---|---|
| ⟨文表層辞書⟩ | |
| ⟨文表層辞書項目⟩... | |
| ⟨表記情報⟩ | :文の表記を示す⟨文表記⟩を定義する。句、中核命題、文型等も含む。 |
| ⟨構造情報⟩ | |
| ⟨要素カテゴリ⟩ | :句、文等の統語範疇を示す⟨文タイプ⟩を定義する。 |
| ⟨構成構造⟩ | 文構成要素への分割情報である⟨形態素列⟩、文の統語構造を木で表示する⟨構文木⟩を定義する。 |
| ⟨指示情報⟩ | :文義の同一性を示す識別記号である⟨文義識別子⟩、同義となる単文の文章などを示す⟨文義⟩、文義が対応する文章表層辞書項目を表示する⟨文脈⟩、パラフレーズされた文を表示する⟨言語内指示⟩、アライメント情報を含むパラレルコーパスに対応する⟨言語間指示⟩を定義する。 |
| ⟨環境情報⟩ | |

## 2.4 文章表層辞書

文章では文間の相互参照がインプリシットに行なわれる。

| | |
|---|---|
| ⟨文章表層辞書⟩ | |
| ⟨文章表層辞書項目⟩... | |
| ⟨表記情報⟩ | :文章の表記を示す⟨文章表記⟩を定義する。 |
| ⟨構造情報⟩ | |
| ⟨要素カテゴリ⟩ | :評論、報道記事など文章の種別を分節した⟨文章タイプ⟩を定義する。 |
| ⟨構成構造⟩ | :文章論的な文間構造を明示した⟨文章構造⟩を定義する。 |
| ⟨指示情報⟩ | :同義となる要約文、要約文章を示す⟨文章義⟩、出典となった文書表層辞書項目を示す⟨文脈⟩、パラフレーズされた文章を表示する⟨言語内指示⟩、文章のパラレルコーパスに対応する⟨言語間指示⟩を定義する。 |
| ⟨環境情報⟩ | |

## 2.5 文書表層辞書

| | |
|---|---|
| ⟨文書表層辞書⟩ | |
| ⟨文書表層辞書項目⟩... | |
| ⟨表記情報⟩ | :文書の表記を示す⟨文書表記⟩を定義する。 |
| ⟨構造情報⟩ | |
| ⟨要素カテゴリ⟩ | :論文、帳票等の文書の種別を示す⟨文書タイプ⟩を定義する。 |
| ⟨構成構造⟩ | :マークアップ言語による構造記述である⟨文書構造⟩を定義する。 |
| ⟨指示情報⟩ | :⟨文書義要約文⟩と⟨文書ストーリ構造⟩により定義される⟨文書義⟩、翻案等を示す⟨言語内指示⟩、翻訳文書である⟨言語間指示⟩を定義する。 |
| ⟨環境情報⟩ | |

## 3 EDR電子化辞書における実現

EDR電子化辞書では、5種の辞書によって表層辞書を実現する [2]。言語の種類は、日本語と英語の2種である。表層辞書との対応は以下の通り。

| 語表層辞書 | 文表層辞書 | 文章・文書表層辞書 |
|---|---|---|
| 単語辞書 | コーパス | テキストベース |
| 対訳辞書 | 共起辞書 | |

**単語辞書** 語表層辞書に対応する。語表層辞書の主要部分を担い、その要求仕様をほぼ実現している。表記情報としては、表記、不変化部指定、発音が定義される。構造情報としては、左接続素性と右接続素性の対 (それぞれ約100種) により接続情報を実現し、また、品詞 (約40に分類)、表層格、アスペクト等の構文情報も定義される。指示情報としては、全内容語に語義説明文および語概念表記が定義される。環境情報として、位相、語種、頻度が定義される。

**対訳辞書** 語表層辞書において、指示情報のうち言語間指示を与える役割を持つ。訳語は必ずしも源言語と同義であるとは限らないので、訳語種別等の情報も定義される。

**コーパス** 文表層辞書に対応する。ただし、単言語コーパスであるため、指示情報の充実を図る必要がある。

**共起辞書** 句を対象とした文表層辞書に対応する。実際には、コーパスから、語の共起に関する情報のみ抽出しコンパクトにまとめたものである。

**テキストベース** 文章・文書表層辞書に対応するが、実現されているのは表記情報と簡単な構造情報のみである。

各辞書の実現規模を以下に示す。

| 種別 | 規模 (日本語) | 規模 (英語) |
|---|---|---|
| 単語辞書 | 25万語,20万概念 | 20万語,26万概念 |
| 対訳辞書 | 25万語,60万訳語 | 20万語,42万訳語 |
| 共起辞書 | 115万タプル | 60万タプル |
| コーパス | 28万文,150万語 | 22万文,100万語 |
| テキストベース | 2000万文 | 500万文 |

## 4 おわりに

汎電子化辞書の表層レベルの構造およびEDR電子化辞書による実現状況を述べた。EDRにおいては、語表層辞書の要求仕様をほぼ満たしたが、文表層辞書については、まだ中途の段階にある。また、文章・文書表層辞書については、電子化テキストを収集しただけの段階である。今後は、語表層辞書の分野の拡大を図ると共に、文表層辞書の充実に注力すべきものと考える。

## 参考文献

[1] 横井、安原、村木、原田、丸山：”汎電子化辞書：言語知識のアーキテクチャ”, 言語処理学会全国大会 B3-2 (1995).

[2] 日本電子化辞書研究所：”EDR電子化辞書マニュアル”, http://www.iijnet.or.jp/edr (1995).