

## 概念レベルにおける電子化辞書の情報構造

横井俊夫<sup>†1,☆</sup> 仲尾由雄<sup>†2,☆</sup>  
 荻野孝野<sup>†3</sup> 田中裕一<sup>†4,☆</sup>

大規模な電子化辞書（その情報内容である言語知識）が概念レベルで持つべき情報構造を明らかにする。ここでいう電子化辞書とは、通常の辞書ばかりではなく、シソーラス、コーパス、テキストペースなどを含む統合的な言語情報（言語知識）のことである。概念レベルは意味を扱う深層レベルの中で基準となる役割を果たす。表層レベルに最も近く、それに沿う情報構造を持つ。なお、この情報構造はEDR電子化辞書の成果を再整理することにより得られたものである。実現事例としてEDR電子化辞書の概念対応部分を仕様と統計データの両面から説明する。大規模知識ベースなどの議論に見られるように、大規模な情報や知識の構造を解明していく研究の重要性が指摘され始めている。本稿の内容は、本格的な実現事例を持つ初めての試みとなっている。

### The Information Structure of Electronic Dictionaries at the Concept Level

TOSHIO YOKOI,<sup>†1,☆</sup> YOSHIO NAKAO,<sup>†2,☆</sup> TAKANO OGINO<sup>†3</sup>  
 and YUICHI TANAKA<sup>†4,☆</sup>

This paper describes a model of the information structure of large-scale electronic dictionaries at the concept level that contain wide-ranging linguistic knowledge. The term *electronic dictionary* in this paper means an integrated body of linguistic information and knowledge that includes the information provided by thesauri, tagged corpora, and raw corpora as well as ordinary dictionaries. The concept level plays an important role for deep levels containing the information of semantic processing. It is the nearest to the surface level and its structure is similar. This information structure is obtained by rearranging the structure of the EDR Electronic Dictionary. An example of actual realization of the information structure is described in view of both the specifications and numeric data of the EDR Dictionary at the concept level. Recently, the importance of the research on the structure of large-scale information and knowledge has become a focus of interest, as shown in the discussions for large-scale knowledge bases, etc. This paper introduces the results of the first trial including full-scale example of actual realization.

#### 1. はじめに

大規模な言語知識が概念レベルで持つべき情報構造<sup>☆☆</sup>を明らかにする。言語知識とは、言語の使い方に関する知識のことであり、具体化されたものを電子化辞書と呼ぶことにする。すなわち、電子化辞書の情報

内容が言語知識である。

言語知識は、言語表現（言語で表現されたもの）を表記のまま直接扱う表層レベルから、表記の表す意味を細分・整理して高度に扱う深層レベルに分けられる。深層レベルではさまざまな意味表現形式による記述が可能で、それぞれの表現形式に対して異なる深層レベルが設定される。その中で、最も表層に近く、それだけに高い一般性を持つレベルが概念レベルと呼ばれるものである。

ここでは、2つの側面から概念<sup>☆☆☆</sup>という言葉を用

†1 電子技術総合研究所知能情報部  
 Machine Understanding Division, Electrotechnical Laboratory

†2 株式会社富士通研究所メディア統合研究部  
 Media Integration Laboratory, Fujitsu Laboratories Ltd.

†3 株式会社日本電子化辞書研究所  
 Japan Electronic Dictionary Institute, Ltd.

†4 株式会社ジャストシステム東京研究所  
 Justsystem Corporation

☆ 本論文の内容は著者らが日本電子化辞書研究所（EDR）に所属していたときの成果に関するものである。

☆☆ 情報が内部に持つ論理的な構造である。どう検索するか、どうコンパクトに表現するかなどの実際の情報の扱いに関するこことは含めない。

☆☆☆ 概念の定義は3章の3.1節、3.2節、3.3節において順次行われる。それまでは、概念という言葉で想起される一般的な意味内容を前提として、輪郭の説明を行う。

いる。1つは、言語によって表現される情報や知識（対象知識、世界知識）を概念のレベルで扱うということである。すなわち、知識のレベルで扱うというほどには深い意味の扱いには立ち入らないということである。もう1つは、個別の意味の表出を包括し、抽象化した意味内容を扱うということである<sup>☆</sup>。言語表現は、具体的な文脈の中でさまざまな意味と対応付けられるが、それらを包括する形、すなわち、概念化された形の意味を基本的な対象とする。

表層レベルでは、語、句、文、文章、文書という言語表現の構成単位に沿って言語知識が記述される。表層に近いということから、概念レベルにおいてもこの構成単位に沿うことになる。したがって、概念レベルは、表層レベルに対応した情報構造を持つ。深層レベルの中で概念レベルを1つの基準となるものとして設定する理由は以下の4つである。

- (1) 言語学においても、自然言語処理においても、類似の考え方に基づく研究がなされ、多くの蓄積がある。フィルモアの格文法理論やシャンクの概念依存関係などから始まった手法の延長上に概念レベルの考え方は位置している。
- (2) 言葉の意味を表層に近い浅いレベルで扱うため一般性が高い。また、表層表現の多義性を概念レベルの意味記述の単位として整理することで、表層レベルでの記述より正規化した形で言語知識を体系化することができる。概念レベルにおける正規化の度合いは初段階のものであるが、このレベルの意味処理だけでも、有用な自然言語処理機能を実現することができる。たとえば、いろいろな曖昧性解消機能などである。
- (3) より深い意味の表現形式のレベルに対し、表層レベルからの変換プロセスの適切な中間ステップとなる。概念レベルを経由することによって他の深層レベルの開発が容易になる。
- (4) 表層に近いということから、表層レベルを利用して本格的な開発が可能である。

最近、広く共有される大規模な知識ベースの研究開発や整備の重要性が指摘されるようになってきている<sup>1)</sup>。いろいろなアプローチの中で、言語知識からのアプローチ、あるいは自然言語を仲介言語とするアプローチの有望性が主張されている<sup>2)</sup>。概念レベルの言語知識は、一般的な知識に対する大規模知識ベースの基底となる部分に対応付けることができる。すなわち、

概念レベルの情報構造は大規模知識ベースの情報構造や知識構造を明らかにしていくうえでの最初の大きな手掛かりとなる。ただし、いかなる知識に対するものであれ、大規模な情報構造や知識構造を明らかにするという本格的な研究は始まったばかりである。したがって本稿の議論も情報構造の基本仕様、概念仕様に焦点をしぼったものとなっている。

以下、2章では、言語知識の情報構造の全体を概観し、概念レベルの位置付けを明らかにする。3章では、概念レベルの言語知識、すなわち電子化辞書の概念辞書部分の情報構造の基本仕様を詳述する。そして、その基本仕様の実現事例としてEDR電子化辞書の対応部分を4章に説明する。ただし、EDR電子化辞書はあくまでも一部の第一ステップとしての実現である。5章では、国内外の類似の研究開発との比較対応を説明する。

## 2. 全体構造と概念レベル

言語知識の情報構造の骨格となる全体構造を説明し、概念レベルの位置付けを明らかにする。以下の説明では理解しやすさを考慮し、言語知識という言葉より電子化辞書という言葉を用いる。ここでいう電子化辞書とは、通常の辞書ばかりではなく、シソーラスやコーパスやテキストデータなどを含む統合的な言語データのことである。汎電子化辞書という言葉も用いられる<sup>3)</sup>。

電子化辞書は、記述の単位、記述のレベル、言語の種類の3点で特徴付けられるサブ辞書群によって構成される（図1）。

### (1) 記述のレベル

言語表現のどのレベルの知識を対象にするのかである。表層レベル<sup>4)</sup>から意味記述にかかる

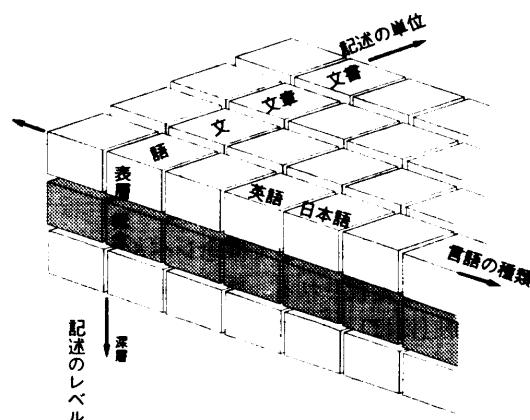


図1 全体構造と概念レベル

Fig. 1 Whole structure and concept level structure.

<sup>☆</sup> 通常の辞書における語義も基本的に同じ考え方に基づくが、本稿でいう概念は、この考え方を敷衍・徹底したものである。

多様な深層レベルまでがある。深層には意味の表現形式に対応していくつものレベルが設定される。意味表現に対してもいろいろな提案がなされてきたが、実際には部分的な適用であったり、理論的な枠組みの提案であったりで、大規模な電子化辞書に適用できる段階ではない。その中で、概念レベルが、表層レベルに直接的な対応を持つ基準となるものとして設定される。

## (2) 記述の単位

言語表現のどのような構成単位を対象にしているかである。語、句、文、文章、文書という構成単位を基本とする。概念レベルもこの構成単位に対応した構造を持つ。

## (3) 言語の種類

どの言語の言語表現を対象にしているのかである。基本的な情報構造は言語に共通である。

言語知識を表現するための基本となる情報単位が辞書項目である。概念レベルでは概念辞書項目である。概念レベルのサブ辞書は概念辞書項目の集合として構成される。概念辞書項目の基本部分は、構成単位や言語の種類によらず、すべてのサブ辞書に共通である。概念辞書項目には自身の情報や他の辞書項目への関係情報が記述されている。この関係情報によって辞書項目間の対応付けが表現され、情報構造が形作られる。関係には、サブ辞書間にまたがる辞書間関係とサブ辞書内の辞書内関係がある。主要な辞書間関係は次の3つである。

### (a) 構成関係

記述の単位の軸に沿って概念レベルの隣り合ったサブ辞書間に定義される。大きな構成単位に対してどのような要素として働くのか、あるいは小さな構成単位のどのような文脈となっているのかを示す。

### (b) 対意関係

記述のレベルの軸に沿って異なった記述レベルのサブ辞書間に定義される。たとえば、表層レベルと概念レベルの間に定義され、言語の表層表現がどの概念に対応するのかを示す。

### (c) 対訳関係

言語の種類の軸に沿って概念レベルのすべてのサブ辞書間に定義される。異なる言語の辞書項目どうしが（ほぼ）同義であるという対応を示す。

\* 表層の表現の構造とその概念の表現の構造とが、要素間の対応、構造上の対応、ともに直接的で単純に対応付く。述語論理、モンタギュー文法、状況意味論のように高度な推論機能を前提とする意味表現を用いる深層レベルとの対応付けはより複雑になる。

概念レベルを含む深層レベルの言語知識の役割は大きく2つ考えられる。1つは、表層表現からそのレベルの意味表現にあるいはその逆に、適切にかつ効率良く変換処理を行うための知識となることである。もう1つは、そのレベルの意味表現の比較処理をして、ある意味的関係が成り立つか否か、あるいはどのような意味的関係が成り立つか、あるいは成り立つ度合いはどのようなものか等々を判断処理するための知識となることである。ここで意味的関係とは、同義、類義、反義、上位、下位などを指す。そして、それぞれの深層レベルの意味表現形式が、どのくらい複雑なものに対し、また、どのくらい広範囲のものに対して上記のような知識を与えられるかによって、そのレベルで取り扱うべき言語知識の種別が決まる。概念レベルは、表層表現に密に対応したレベルであるので、言葉で陽に表現される部分を扱う。会話のようにその意味が状況に強く依存するものについては、別の仕組みが必要である。概念レベルは、あまり複雑でないものに対し、そして、かなり広範囲のものに対して、意味処理のための知識を与える。読み手に対して特殊な了解事項を要求しない記述的な文書、すなわち、事実や自己の見解などを広く伝えるために書かれた記事や論文などを扱うための基礎的な言語知識を与える。このように限定することによって、各種の曖昧性の解消機能や、検索、翻訳、要約などのより高度な文書処理機能を実現するのが概念レベルの役割である。

## 3. 概念辞書の情報構造

概念レベルのサブ辞書の情報構造を詳細化する。語概念辞書、文概念辞書、文章概念辞書、文書概念辞書の情報構造\*\*である。ただし、文章概念、文書概念について、大枠としての説明にとどめる。

### 3.1 概念と概念辞書項目

概念と称するものの基本的な定義と概念辞書項目の概略仕様を説明する。続く各節で構成単位ごとの概念と辞書項目の詳細化が行われる。

**概念：** 概念は実体概念と関係概念に分けられる。実体概念とは、もの、こと、事象、事象列など実体を有するものに対応する概念である。関係概念は実体概念どうしのかかわり方を表す概念である。1つの概念を実体概念と見ると、関係概念と見るのかは、観点によって恣意的となる側面がある。概念辞書では、表層表記に直接的に対応付けられ

\*\* 句概念辞書は、語概念辞書と文概念辞書の両者の性質を持つ。ここでは説明を省略する。

<概念辞書項目>	
<識別情報>	: 概念の識別や基本的な定義
<概念構成情報>	: 概念の構成構造
<概念関係情報>	: 他の辞書項目との意味的な関係
<階層関係情報>	: 概念辞書項目間の階層関係
.....	: 概念辞書項目間のその他の意味的な関係
<構成関係情報>	: 上位、下位の構成単位の概念辞書項目との関係
<対意関係情報>	: 上層、下層の記述レベルの概念辞書項目との関係
<対訳関係情報>	: 他言語の概念辞書項目との関係

図2 概念辞書項目

Fig. 2 Concept dictionary entry.

るようにするという観点から概念の種別が決定される。さらに、語、文、文章、文書という表層での記述単位（構成単位）に対応して、それぞれ実体概念と関係概念が定義されることになる。

**概念辞書項目：**1つ1つの概念に概念辞書項目が対応付けられる。図2に（実体概念に関する）概念辞書項目の基本構造を示す。概念を識別するための情報と基本的な定義情報が<識別情報>に記述される。構成要素となる概念からその概念がどのように構成されているのかなどの情報が<概念構成情報>に記述される。概念は、他の概念との意味的な関係を列挙することによってさらに詳細な定義がなされる。<概念関係情報>には主要なものとして4つの意味的な関係情報があげられる。構成関係、対意関係、対訳関係は2章で述べた辞書間関係である。辞書内関係である階層関係は属種関係とも呼ばれるもので、概念の上位－下位関係を表す。階層関係は概念レベルの意味表現の比較処理において判断の基礎となる知識である。ただし、現在のところ明確な意味付けを持つのは語概念辞書と文概念辞書である。

このような概念レベルの情報構造によって、語概念から構成関係をたどって、小さな構成単位の概念から大きな構成単位の概念がどう構成されていくのかを段階的に記述することができる。逆に、文書概念から概念構成情報をたどることによってマクロな構造を記述することができる。そして、各々の構成単位に対して類似性を判断する知識を階層関係によって記述することで、個別には記述しきれない多様な言語表現に対して、辞書に記述された知識を援用する手段を与えることができる。つまり、各々の構成単位で記述された階層関係の知識と、構成関係情報・概念構成情報とを合わせて用いることで、構成単位自身の類似性、その構成単位が置かれた文脈・発話状況の類似性、その構成単位を構成する下位の要素および構造の類似性、というように、いくつもの観点から言語知識を利用する手

段が与えられる。このようにして、広範囲の処理に必要な知識を統一的に整理することができる。

### 3.2 語概念辞書

表層レベルで語と称されているものが表す語概念と、その内容記述を行う語概念辞書項目について述べる。

#### 3.2.1 語概念

名詞、動詞、形容詞、副詞などの概念語（内容語）と通常呼ばれているものによって表されるものが実体概念である。一方、助詞、前置詞などの関係語と通常呼ばれているものによって表されるものが関係概念である。関係概念は語の位置などによって表される場合もある。この場合は、関係概念は、対応する表層の語を持たないことになる。実体概念についても、また関係概念についても言語学的な観点からの品詞と固定的に対応させるわけではなく、自然言語処理の観点からの実際的な扱いがなされる。すなわち、通常の語と同じような働きを示す慣用的に用いられる表現などは相当語として扱い、それらの働きに応じて1つの概念を対応させる。

語は実際の文、文章の中で、1つの表層表記のままで実際に多種多様な意味を表現する。電子化辞書としては、意味の記述単位である概念として何をもつてくるのか、いかなる観点に立つか、いかなる粒度のものとして見るか、などについての明確な指針が必要である。指針を以下の3点に整理する。まず、実体概念を中心にしてである。

##### (1) 概念化されたもの

実際の文、文章中においては、語はほとんど個別の対象（「花<ある花>を見た.」）を表すが、それらのすべてに共通の性質をまとめ概念化した内容（<花というもの>）を対象とする。いうなれば、クラスを定義するものとして概念を扱う。

##### (2) 慣用化されたもの

実際の文、文章中においては、語はさまざまナレトリカルな意味に対応付けられることがある。

<語概念辞書項目>	
<識別情報>	: 語概念の識別標識
<概念ラベル>	: 自然言語による概念の説明
<概念説明>	: 要素概念による語概念の構成構造
<概念構成情報>	: 意味素性による記述
<意味素性構造>	: その他の要素概念による語概念の構成構造の記述
.....	
<概念関係情報>	
<階層関係情報>	: 語概念の属種関係情報
<全体一部分関係情報>	: 全体となる語概念とその部分となる語概念との関係情報
<反義関係情報>	: 反義となる語概念どうしの関係情報
.....	: 語概念どうしのその他の意味的な関係
<構成関係情報>	: この語概念を要素とする文概念への関係情報
<対意関係情報>	: 表層やより深層の語辞書項目への関係情報
<対訳関係情報>	: 他言語の語概念との対訳関係情報
<言語1対訳>	: 言語1の語概念との対訳関係情報
<言語2対訳>	: 言語2の語概念との対訳関係情報
.....	: その他の言語の語概念との対訳関係情報

図3 語概念辞書項目

Fig. 3 Word-concept dictionary entry.

この場合で慣用化し、定着している場合は、派生義（「花<桜の花>」）を原義（「花<植物の花と呼ばれる器官>」）とは独立した概念として扱う。

### (3) 一体化されたもの

多くの属性、あるいは多くの要素概念を含む一体化されたものとして扱い、属性（要素概念）への分解は別途適切な手段を講ずる（階層関係による多重継承や事例による推論の機構など）。実際の文、文章中においては、それぞれの属性（要素概念）が個別に対応付けられる場合が多くある（「学校<学校という建物>が見える。」「学校<学校という場所>まで歩く。」「学校<学校という組織>が判断した。」）。しかし、個別の属性（要素概念）に分解する作業をこの段階では行わない（「学校<学校という教育施設>」）。

関係概念は、実体概念ほどバラエティに富むものではないが、いろいろな詳細化の段階が設定できる。しかし、まずは荒い近似の段階で電子化辞書の開発を行い、利用経験を蓄積する中で詳細化を進めるのが妥当である。

次に、表層語と語概念との対応付けを行い、概念の定義を実質的なものにする。表層レベルにおいて、各単語には語義が対応付けられている<sup>4)</sup>。多義語には複数の語義が対応している。まずこの語義を上記の指針に従って整理する。単語ごとに整理された語義を、異なる単語の間で比較して、同義と見なされる語義どうしを統合化する。このようにして得られたものを語概念とする。

### 3.2.2 語概念辞書項目

実体概念に対する語概念辞書項目を図3に示す。<識別情報>は辞書を利用する人間ないし計算機が語概念を識別するための情報である。<概念ラベル>は語概念を識別するための標識であり、概念内容を容易に想起できるように工夫されたものである。対応する表層語を利用するために語義ラベル（表層語+語義番号）を用いる。複数の語義が統合化されている場合には、語義ラベルの集合か代表語義ラベルを用いる。<概念説明>は自然言語文による概念の説明である。形式などは語義説明文に準ずる。概念ラベルを補足して、概念内容を正確かつ容易に人間に理解させるのが役割である。

<概念構成情報>はこの語概念を要素概念に分解したときの構造である。どのようなものに分解するかについては、さまざまな立場からさまざまな提案がなされている。意味素性構造、コーパスや定義文における他の単語との関連を表すべきトルなどがある。

<概念関係情報>は辞書内関係と辞書間関係にかかる情報の記述である。辞書内関係には階層（属種）関係、全体一部分関係、反義関係などの意味的関係が含まれる。これらによって同一サブ辞書内の語概念が関係付けられる。これらの関係の中で最も重要なものが階層関係である。その語概念の上位となる語概念（複数の可能性あり）と下位となる語概念が対応付けられる。この上位、下位の判断を行う観点を何にするかによって異なる階層関係が定義される。自然言語処理の立場から見ると、述語概念の選択制約の処理な

<文概念辞書項目>	
<識別情報>	
<文概念ラベル>	: 表層文を用いた文概念の識別標識
<文概念説明>	: 文、文章による説明やバラフレーズ
<概念構成情報>	
<文概念構造>	: 語概念などを要素概念とした文概念の構造記述
<概念関係情報>	
<同義関係情報>	: 同義となる文概念との対応情報
<階層関係情報>	: 上位-下位となる文概念との対応情報
.....	: 文概念どうしのその他の意味的な対応情報
<構成関係情報>	: 文章概念への対応情報
<対意関係情報>	: 表層やより深層の文辞書項目への対応情報
<対訳関係情報>	: 他言語の文概念との対訳対応情報
<言語1対訳>	: 言語1の文概念との対訳対応情報
<言語2対訳>	: 言語2の文概念との対訳対応情報
.....	: その他の言語の文概念との対訳対応情報

図4 文概念辞書項目  
Fig. 4 Sentence-concept dictionary entry.

のために、文概念の中での共起の仕方と強く結び付いた観点による階層関係が重要となる。語概念Aを含む文概念を例挙し、各文概念中の語概念Aを語概念Bで置き換えるてもすべてのものが文概念として成り立つとき、語概念Aを語概念Bの上位概念とするという観点である<sup>5</sup>。階層関係によって語概念の体系が形成されることになる。なお、この体系の上位には、体系構成上の必要性から、対応する表層語を持たない仮想された語概念が配置される。

辞書間関係に基づく<構成関係情報>は、この語概念を要素として含む文概念（隣接する文概念辞書内の文概念辞書項目）すべてへの対応付けである。<対意関係情報>は、語概念辞書項目から語表層辞書項目への対応関係を与えるもので、語概念からそれを表しめる表層の語すべてへの対応付けを行うものである。また、より深層での意味記述がなされている場合は、その辞書項目への対応付けもなされる。<対訳関係情報>は、他のすべての言語の語概念で同義となるものへの対応付けである。同義となるものがない場合は、一番近い上位概念への対応付けがなされる。

辞書項目内の各サブ項目も互いに関係し合う。たとえば、<概念説明>、<概念構成情報>、<階層関係情報>は語概念の内部構造の詳細化という共通の性質を持つ。このような性質は、電子化辞書の開発工程で有効利用できる。

### 3.3 文概念辞書

表層レベルで文と称されるものが表す文概念とその内容記述を行う文概念辞書項目を説明する。文は語に

比べ格段と多様になる。対象とする文の種類を何にするかによって文辞書の性格が異なってくる。自然言語処理の立場からは単文を網羅し複文は文章辞書の中で扱うという考え方もある。

#### 3.3.1 文概念

文の表す事象を中心とした内容が実体概念である。文の並び具合、接続詞などで関係概念が表される。文の実体概念を定義していくにあたっては、文概念の内部構造をどのようなものとして見るのかが重要な判断となる。自然言語処理の立場からは、次のような見方が妥当である。すなわち、その事象に何がどのようにかかわっているのかを示す中核となる命題に、相、時制、極性、モダリティなどの属性を付加した構造を持つというものである<sup>5)</sup>。そこで文（実体）概念は、命題、相、時制、極性、モダリティなどの属性を付加した構造を持つ、あるいはそのような要素概念から構成される概念として定義されることになる。そして命題部分は深層格パターンの考え方にならい、述語概念（語実体概念）を中心に修飾する語の実体概念と修飾の仕方を表す関係概念によって構成する。

表層文と文概念の対応付けは単純である。すなわち、表層レベルで表層文に割り振られた文義（文義ラベル）をそのまま文概念（文概念ラベル）に対応付ける。同義となる文概念どうしの対応付けは、陽に記述される。対応付けの仕方を陽に知ることが重要だからである。

#### 3.3.2 文概念辞書項目

文概念辞書項目を図4に示す。

<識別情報>は辞書を利用する人間ないし計算機が文概念を識別するための情報である。<文概念ラベル>は、対応する表層文をほぼそのまま文概念の識別標識としたものである。<文概念説明>は表層文を可

\* これは、この観点の定義上のことであり、実際の階層関係の判断プロセスにこのまま用いられるわけではない。

読性の高い、ないしは構造の明確な文、ないしは文章へ言い換えて、意味を明解に把握できるようにしたものである。

<文概念構造>は文概念を命題、相、時制、極性、モダリティなどの要素概念の合成構造として表現したものである。命題を表現する語概念は文という文脈では、通常はあるインスタンス（あるモデルにおける対象）を表す。ただし荒い近似として語概念のまま用いた命題の表現も利用される。この場合、命題は格パターンに対応する。

<概念関係情報>の辞書内関係には同義関係、階層関係などがある。<同義関係情報>は、同義となる文概念の対応付けである。対応付けは文概念構造の内部構造を対応付ける仕方で行われる。<階層関係情報>は、文概念の上位-下位の関係である。この関係も語概念の階層関係と同様に観点によって多数のものが考えられる。その中で、表層に近く、したがって安定した判断ができ、また、辞書の構成上有用なものとして、次のような観点がある。命題、相、時制、極性、モダリティなどを文の要素概念（概念属性）としたときに、これがより詳細に特性付けられているものを下位とするという観点である。命題に関しては、構成要素となる語概念が詳細に規定されるほど下位とするという観点である。この観点は<文概念構造>に直接対応する。

辞書間関係に基づく<構成関係情報>はこの文概念を要素として含む文章概念への対応付けである。<対意関係情報>は、この文概念と表層文との内部構造を含めた対応付けである。また、より深層での意味記述がなされている場合は、それへの対応付けもなされる。<対訳関係情報>は他のすべての言語の文概念で同義となるものへの文概念構造の上の対応付けである。

### 3.4 文章概念辞書・文書概念辞書

文章と文書に関する概念辞書は、大枠としては本稿の情報構造に収まるものの、細部にわたっては、まだ、つめるべき課題が多い。ここでは文章概念に関する概要だけの説明にとどめる。ここでいう文章とは、複文、段落、段落列を含む。文書とは、それ単独でまとまりのある体系的な情報を表現するもので、論文、記事などさまざまな形態をとる。

文章の概念構造を考える土台となるのは、談話分析や物語分析における研究成果である。この成果を敷衍すれば、文章概念は、文が表す事象概念を実体概念とし、時間的前後関係や因果関係などを関係概念として表される事象列の構造を持つことになる。そして、文書概念の構成要素として文書の全体構造の中に位置付けられることになる。これにより、部分構造における

結束性と全体（マクロ）構造における整合性の両面から言語知識が表現できる。すなわち、事象の主題、生起した時点（テンス）、生起の仕方（アスペクト）、話者の意図（モダリティ）等々の事象概念を特徴付ける属性の連鎖の仕方や、説明文における導入・展開・評価・結論や物語における背景・挿話（できごと・反応の組）といった文章構成の類型などの知識である。

ただし、これらに関する研究成果は、文章や文書に関する言語知識を体系的に整理するための枠組みとしては、まだ十分とはいえない。これから重要な研究課題である。

## 4. EDR電子化辞書における実現

前章で述べた概念辞書の情報構造の実現事例としてEDR電子化辞書<sup>6)</sup>の概念辞書対応部分を説明する。これによって情報構造の妥当性を実証するとともに、情報構造の具体的な実現の方法が明らかにされることになる。辞書仕様と統計データの両面から実現の有り様を説明する。

### 4.1 辞書仕様

情報構造は、電子化辞書が全体として持つべき情報の種類と相互の関係という立場で仕様を検討したものである。一方、EDR電子化辞書は辞書を利用者に提供するに際しての現実的な要請を加味して仕様が決定されている。そのために生じた構造上のずれを補正しつつ説明する。また、EDR電子化辞書は日本語と英語に対し、日常一般に用いられる基本語と情報処理用語を対象にしているが、ここでは基本語に由来する部分に全体を代表させる。

語概念辞書の基本部分の実現となるのがEDR電子化辞書の概念見出し辞書と概念体系辞書である。文概念辞書の部分的な実現になるのが概念記述辞書、共起辞書、EDRコーパスである。文章・文書辞書に相当するものとしてはEDRテキストベースがあるが、これには概念レベルに相当する情報は付与されておらず、概念レベルの実現は存在しない。なお、EDR電子化辞書では、概念見出し辞書、概念体系辞書、概念記述辞書を合わせたものを概念辞書と呼んでいる。これは利用者の便宜を考慮してそのようにまとめたものである。

辞書項目に対応するのが辞書レコードである。図5に概念見出し辞書と概念体系辞書の辞書レコードの構造を示す。<概念見出し辞書レコード>の内容が語概念辞書の<識別情報>に対応する。<階層関係情報>を上位概念と下位概念の対に分解したものが概念体系レコードである。これは管理の便宜上の処置である。階層関係によって構成される全体の構造物を概念体系

と呼んでいる。その上位層部を図 6a) に、体系をたどった例を図 6b) に示す。a) の概念名で単語以外の

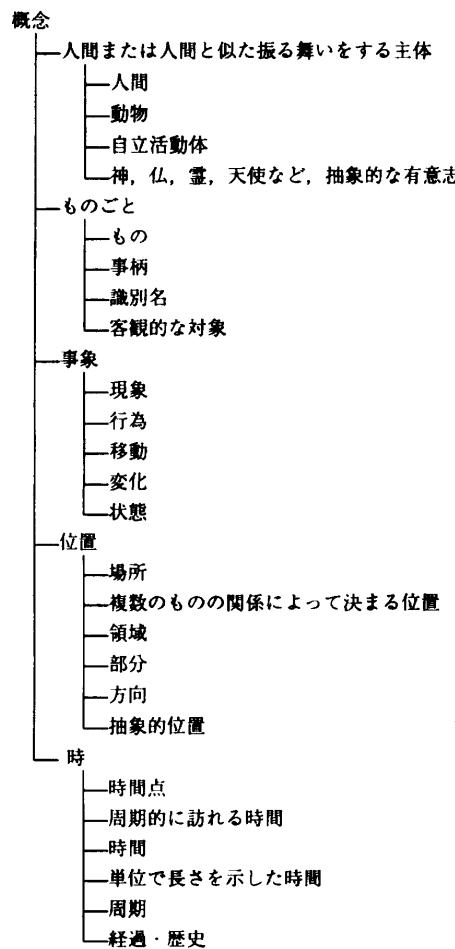
<概念見出し辞書レコード>	
<レコード番号>	: レコードタイプと識別番号
<概念識別子>	: 概念を一意に特定する整数
<概念見出し>	: 概念に意味を表現する単語
<概念説明>	: 概念の内容を表現する説明文
<英語概念説明>	: 英文による概念の説明
<日本語概念説明>	: 和文による概念の説明
<管理情報>	: レコードの管理履歴情報

<概念体系辞書レコード>	
<レコード番号>	: レコードタイプと識別番号
<上位概念識別子>	: 上位となる概念の識別子
<下位概念識別子>	: 下位となる概念の識別子
<管理情報>	: レコードの管理履歴情報

図 5 EDR 電子化辞書の語概念辞書対応部分

Fig. 5 Records of the EDR electronic dictionary corresponding to the word-concept dictionary.



a) 基本語概念体系の上位層部  
Top-level structure of the EDR concept classification.

形式のものは、対応する表層語を持っていない。<構成関係情報>は検索によって得られるため陽には示されていない。<対訳関係情報>には、日本語由来の概念と英語由来の概念と一緒にして、同義や階層関係の処理を行うことによって対応している。

文概念辞書に対応する部分を含む辞書レコードを図 7 に示す。EDR コーパスでは、一文が一文義、一文概念である。<EDR コーパスレコード>には表層レベルの情報も合わせて記述されている。したがって <文概念ラベル> は省略されている。<意味情報>が文概念辞書の <文概念構造> に対応する。これは、命題を文中の自立語の概念と概念関係子によるネットワークで表現し、相、時制、極性、モダリティ（発話の意図、話者の視点、文体など）を属性子や補助的な構造を附加して表現したもので、機械翻訳における中間言語としても使用可能なものである<sup>7)</sup>。文概念の命

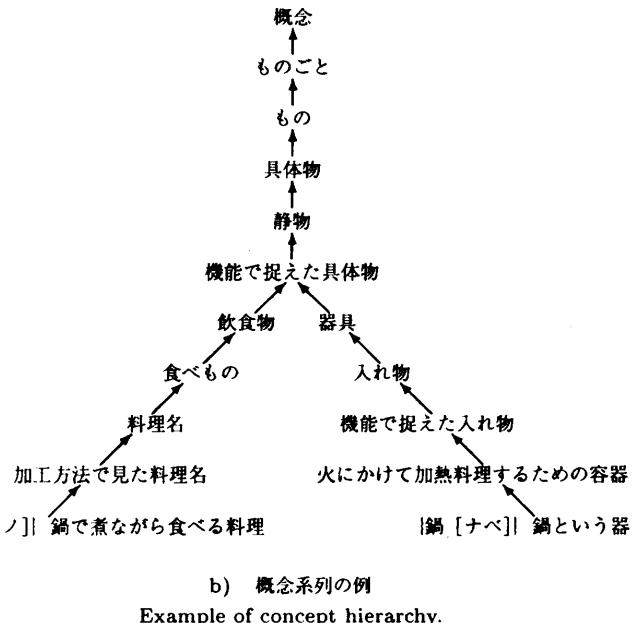


図 6 基本語概念体系  
Fig. 6 The EDR concept classification of general vocabulary.

<EDR コーパスレコード>	
<レコード番号>	: レコードタイプと識別番号
<文情報>	: 文識別子と文見出し
<構成要素情報>	: 文を構成する形態素列
<形態素情報>	: 形態素、複合語の構成情報
<構文情報>	: 構文木
<意味情報>	: 文概念の構成構造（中間言語表現）
<管理情報>	: レコードの管理履歴情報
<共起辞書レコード>	
<レコード番号>	: レコードタイプと識別番号
<見出し情報>	: 共起句見出し
<共起句構成要素情報>	: 共起句を構成する形態素列
<構文情報>	: 共起句（最簡文）の構成情報
<意味情報>	: 共起句（最簡文）概念の構成情報
<共起状況情報>	: コーパスにおける出現頻度と出現部分の周辺
<管理情報>	: レコードの管理履歴情報
<概念記述辞書レコード>	
<レコード番号>	: レコードタイプと識別番号
<記述区分>	: 最簡命題概念の区分
<記述タブル>	: 最簡命題概念の構成情報
<管理情報>	: レコードの管理履歴情報

図 7 EDR 電子化辞書の文概念辞書対応部分

Fig. 7 Records of the EDR electronic dictionary corresponding to the sentence-concept dictionary.

題部分に対し、階層関係の上位にあるものが<概念記述辞書レコード>の<記述タブル>である。これは、述語となる語概念とそれを修飾するもう1つの語概念とを関係概念で結んだという最簡命題概念である。最簡命題概念は、最も簡単で要素的な文概念（命題）であるが、依存関係の強い述語概念と修飾概念の組合せは、その述語がとりうる全体の格パターンと同等の制約として働きうるものである。この最簡命題概念には、コーパスの文概念に結び付けられたものと、概念体系の上位層で作成した、表層表現を持たないものとに分けられる。この区分が<記述区分>に記されている。表層表現との対応付けを陽に記述したのが<共起辞書レコード>である。ただし、これも最簡文、最簡命題に対するものである。最簡命題は文概念体系上、上位層に位置する。そのため捨象される部分が多くなりすぎるくらいがある。そこで中間層に位置するものとして、日本語の基本動詞に対し設けたものが日本語動詞共起パターン副辞書に記述されている。

#### 4.2 統計データ

語概念辞書の<階層関係情報>に相当する EDR 電子化辞書（第1版）の統計データとして、表1に概念体系の規模に関するデータを、表2に概念体系の全体構造に関するデータを示す。

表1で、「親概念数」とはある概念の直接の上位として関係付けられた概念の異なり数である。親概念数が

表1 基本語概念体系における一概念あたりの親概念数の度数分布

Table 1 Histogram of the number of the immediate super-ordinate concepts related to a concept.

親概念数	概念数	(比率)
1	360,454	(94.6%)
2	19,073	(5.0%)
それ以上	1,608	(0.4%)
合計	381,135	(100.0%)

表2 基本語概念体系における概念の分布

Table 2 Distribution of concepts in the EDR concept classification.

概念項目のグループ	子概念数	(比率)
上層 1~10 項目	4,435	(1.1%)
上層 11~50 項目	11,326	(2.9%)
上層 51~100 項目	17,197	(4.3%)
上層 101~500 項目	105,892	(26.8%)
上層 501~1,000 項目	61,440	(15.5%)
上層 1,001~5,000 項目	167,733	(42.4%)
最下層	27,879	(7.0%)

2以上となっている概念は、いわゆる多重継承の構造に対応するものである。概念体系における多重継承は排他的でない複数の観点からある概念が多重に分類されている場合に生ずる。表1は約5%程度の概念しか多重に分類されていないことを示している。このことは、1つには38万概念という膨大な量の概念を分類していくための工数上の制約から、第一版としては、基本的な部分のみに対応したからである。また、もう1つは語概念の持つ性質のうち文脈によらないもののみを考慮したことにもよる。すなわち、ある文脈において何を指示示すかによって真偽が異なるような性質では分類していない。たとえば（単なる）「皿」という語の概念は、「一般に丸いものである」（形の観点）とか「瀬戸物であることが多い」（材質の観点）などの性質では分類せず、「形のある具体的な物質でできたもの」として分類する。これは、強い連関はあるが実際の真偽が文脈に依存する性質は、事例などから類推するなどの別のメカニズムで取り扱うべきであるとの考え方からである。EDR 電子化辞書では、このような考え方に対しては、概念記述や共起辞書、あるいはコーパスが情報を提供することになる。

表2は概念体系の階層における概念の分布の様子を示している。表2では、まず、類似の概念を下位概念としてまとめている概念、すなわち、概念項目<sup>☆</sup>を、

☆ 通常のシソーラスの分類カテゴリの機能を持つ概念。表層の単語に対応を持つものと対応を持たないものがある。前者は、下位概念を持つという点以外には概念項目でない概念と区別はない。後者は、概念体系を構成するために導入された特別な概念である。

表3 概念記述の規模  
Table 3 Amount of concept descriptions.

関係名	記述タブル		日本語事例数（25万文中）				英語事例数（20万文中）			
	異なり	構成比	延べ	構成比	異なり	重複度	延べ	構成比	異なり	重複度
agent	40,221	7.9%	96,816	10.5%	79,416	1.22	118,808	20.2%	74,248	1.60
object	282,800	55.8%	514,056	55.8%	388,957	1.32	285,263	48.4%	232,038	1.23
a-object	46,161	9.1%	73,602	8.0%	66,036	1.11	64,298	10.9%	57,231	1.12
goal	46,138	9.1%	99,105	10.8%	74,459	1.33	43,453	7.4%	37,699	1.15
implement	20,043	4.0%	25,983	2.8%	24,020	1.08	13,492	2.3%	12,894	1.05
cause	9,623	1.9%	9,547	1.0%	9,067	1.05	5,743	1.0%	5,468	1.05
place	26,039	5.1%	49,519	5.4%	42,833	1.16	29,053	4.9%	25,381	1.14
scene	36,181	7.1%	53,071	5.8%	47,946	1.11	29,407	5.0%	27,823	1.06
小計	507,206	100%	921,699	100%	732,734	1.26	589,517	100%	472,782	1.25
その他	—	—	1,095,061	118.8%	850,091	1.29	577,275	97.9%	477,993	1.21
総計	507,206	100%	2,016,760	218.8%	1,582,825	1.27	1,166,792	197.9%	950,775	1.23

その子孫概念<sup>☆</sup>の多い順に7グループに分け、体系の位置の目安とした<sup>☆☆</sup>。そして、グループごとに、概念項目の直接の下位概念（子概念）の異なり数を集計した。このようにグループ化すると、ある概念がいずれのグループの概念項目の子概念となっているかは、その概念の弁別性の1つの指標になる<sup>8)</sup>。たとえば、上層1,000項目レベルの概念項目の子概念は、概念数全体の1,000分の1のオーダの数の概念群と姉妹関係<sup>☆☆☆</sup>にあることになり、1,000個レベルの異なりを弁別できる程度の情報が与えられていると考えられる。EDR電子化辞書の基本語概念体系中の約半数の概念は、少なくとも1,000項目レベルの弁別性を持っている。50項目レベルより小さい弁別性しか持っていない約4%の概念の多くは、「物事」に対応する概念など指示示す内容自体が広くそれ以上の分類が困難なものである。100項目レベル～1,000項目レベルの弁別性を与えられている概念には、分類の観点を増やせばさらに詳細な分類が可能なものもある。しかしながら、実際に行うには、現在の数千のオーダの分類項目を数万のオーダへと増やすこととなる。この場合、作業者が分類体系の全体を把握することが困難になるので、分類作業の方法を大きく見直す必要がある。

表3に概念記述辞書（第1版）中の「記述タブル」の規模を示す統計データをあげる。これは、文辞書における最簡命題概念に相当する部分である。表中、「記述タブル」とは、概念記述辞書で記述された概念関係の数であり、「事例数」とは、日本語・英語のコーパス

に現れた概念関係の数である。「構成比」は概念記述辞書で記述されている8つの関係の合計数を基準とした百分率で示してある。「重複度」はコーパス中の関係事例の延べ数と異なり数との比である。事例数には、固有名詞に対応する概念など、EDR概念辞書には登録されていない概念の関係事例も含まれている。「その他」の関係とは、名詞と名詞の修飾関係に対応する関係や、述語概念と時間を指定する概念との関係、述語概念どうしの時間的な前後関係など、格バタン以外の関係である。なお、コーパスには実際の文では省略されている要素に対応する概念の関係や文の構造に関する関係なども含まれているが、表3にはこのような関係は含めていない。

### 5. 他の事例との比較検討

情報構造の妥当性の検討をさらに進めるために、国内外の他の代表的な事例との比較検討を行う。ただし、概念レベルに対応する部分についての概観である。事例としては辞書として分類されるものとシソーラスとして分類されるものがある。また、仕様の策定に重点を置き実現の方は実験段階にとどまっているものと、本格的な実現を達成したものとがある。

辞書に分類され仕様策定に重点を置いた事例に欧州のAcquilexとMultilex、日本のIPAL辞書がある。Acquilex<sup>9)</sup>は、機械可読辞書からの語彙知識の半自動獲得技術を目指したものである。目標となる語彙知識ベースの辞書項目は、HPSG流のグラフユニフィケーションに基づく表現言語で記述される。タイプシステムと素性構造の意味構造部分に語概念や文概念を扱う仕組みが設けられている。しかし、辞書としては野心的な形式化を進めたために、本格的な実現を進めるのが難しいものになっている。また、そのような形式化が自然言語処理にもたらす効用の議論も十分とはいえない。本稿の情報構造くらいから段階を追うのが妥当であろう。Multilex<sup>10)</sup>は多言語間の辞書データベース

☆ 概念体系でその概念の下位に位置している全概念。すなわち、直接の下位概念、下位概念の下位概念…をすべて含む。

☆☆ 多重継承を持つ概念体系を正規のN進木に近似し、上層～下層の軸に沿ったゾーンを設定する操作に対応。体系が深さが一定のN進木の形をしている場合は、子孫概念の数が大きい概念ほど体系のルートに近い。

☆☆☆ 同じ概念を直接の上位とする概念どうし（姉妹概念）の関係だけでなく、姉妹概念の下位概念との関係も含めて、ここでは姉妹関係と呼んでいる。

の基準を策定しようというものである。意味情報部分の仕様では語概念辞書の仕様の基本部分に対応するものが検討されている。あくまでも仕様レベルの議論である。IPAL 辞書<sup>11),12)</sup>は統語情報を中心に精度の高い辞書の実現を目指したものである。名詞概念の分類については基本動詞辞書で 18 種類、基本形容詞辞書で 41 種類の意味素性を用いている。意味素性の数は、記述可能な述語の文型パターン・選択制限パターン（本稿の用語でいうと命題のパターン）の種類を決定する。実際の自然言語処理では、何を目的とするかで区別すべきパターンの精密さは異なるが、たとえば日本語に対する英訳語を確定するというプロセスへ適用するためには、このレベルの分解能ではまだ不足であることが報告されている<sup>13)</sup>。また、自然言語処理に適用するためには、述語の選択制限パターンとともに、フィラとなる名詞についての語概念の階層関係などの情報も必要である。IPAL 辞書ではこの部分に対応するものも含めて、名詞辞書の開発も行われている。

自然言語処理、特に機械翻訳のための実用規模の実現を果たしているのが ALT 辞書（機械翻訳システム ALT 用辞書）<sup>13)</sup>である。3,000 の分類項目によって意味属性、すなわち語概念属性の体系が作られ、単語意味辞書の中で各語（語概念）にこれらの属性が割り付けられている。これによって訳語選択レベルでの多義性を十分に解消できる分解能を実現している。構文意味辞書の中に文型という形式で文に対する文概念の扱いが記述されている。

シソーラスとして分類される事例として WordNet, ロジェ・シソーラス、分類語彙表がある。WordNet<sup>14)</sup>は心理言語学の研究のために実現されたもので品詞ごとに意味的な関係が詳しく記述されている。語概念辞書の仕様にはほぼそのまま対応する構造を持っている。自然言語処理のためには表層レベルへの対応付けが必要であるが、その作業も計画されている。ロジェ・シソーラス<sup>15)</sup>は、人間が作文する際に、適切な表現を選び出すことができるようにするためのものである。したがって類義のものを品詞の区別なく集めるという語概念辞書の階層関係とは観点が異なる。その点を除けば、長い歴史に裏打ちされた蓄積として利用できる。分類語彙表<sup>16)</sup>は語彙を意味の世界でとらえようという試みである。多義語の扱いが徹底していない面があるが、やはり大いに利用すべき蓄積である。

以上のように、それぞれの事例は情報構造の部分部分に対応付けられ、全体としてその妥当性を支持するものになっている。

## 6. まとめ

本稿の電子化辞書の情報構造の提案は、このような情報構造を持つ電子化辞書を全体として一挙に実現しようということを目的としたものではない。実際は、それぞれが、必要となる部分を、必要とする規模で開発し利用していくことになる。ここでの提案は、たくさんの努力が収まるべき共通の枠組みを作り、互いに協力し合い、着実に蓄積していくための共有される土台作りへの利用を目的としたものである。

電子化辞書は本来、大規模で、高精度で、低コストで実現できるものでなければならない。そのためには、本稿では十分な余裕がなかったが、検討すべき重要な論点が 2 つある。1 つは、この情報構造がこれからの自然言語処理のトレンドに十分適合できるものであるということに関してである。もう 1 つは、効率の良い実現のプロセスに結び付けることができるものであるということに関してである。特に、これからは、コンピュータによる実現支援技術<sup>17)</sup>の研究が重要で、かつ興味深いものとなる。

本稿のような情報構造の研究を手掛かりとして、これから言語知識や世界知識に対する体系的な研究や技術開発が着実に進展していくことが強く期待される。

謝辞 EDR 電子化辞書プロジェクトに携わった多くの方々に感謝する。

## 参考文献

- 1) Fuchi, K. and Yokoi, T. (Eds.): *Knowledge Building and Knowledge Sharing*, Ohmsha and IOS Press (1994).
- 2) Yokoi, T.: The Impact of the EDR Electronic Dictionary on Very Large Knowledge Bases, *Towards Very Large Knowledge Bases* (Mars, N. (Ed.)), Ohmsha and IOS Press, pp.13-21 (1995).
- 3) 横井俊夫, 安原 宏, 村木一至, 原田千秋, 丸山冬樹: 汎電子化辞書—言語知識のアーキテクチャー, 言語処理学会第 1 回年次大会発表論文集, pp.185-188 (1995).
- 4) 横井俊夫, 木村和広, 小泉敦子, 三吉秀夫: 表層レベルにおける電子化辞書の情報構造, 情報処理学会論文誌, Vol.37, No.3, pp.333-344 (1996).
- 5) 中右 実: 認知意味論の原理, 大修館書店 (1994).
- 6) 日本電子化辞書研究所: EDR 電子化辞書仕様明書(第 2 版), EDR TR-045 (1995).  
[<http://www.iijnet.or.jp/edr> で ftp が可能].
- 7) 國際情報化センター機械翻訳システム研究所: 中間言語(最終版) (1993).
- 8) Nakao, Y. and Ogin, T.: EDR Concept Clas-

- sification and Methodology for Its Development, *Proc. 47th FID Conference and Congress*, pp.122-127 (1994).
- 9) Briscoe, T., et al.: ACQUILEX: Acquisition of Lexical Knowledge for Natural Language Processing Systems, Esprit BRA-3030 Periodic Progress Report, No.1 (1990).
- 10) Serasset, G.: Recent Trends of Electronic Dictionary Research and Development, EDR TM-038, Japan Electronic Dictionary Research Institute (1994).
- 11) 情報処理振興事業協会技術センター：計算機用日本語基本動詞辞書 IPAL (Basic Verbs) —解説編, 61 技-073 (1987).
- 12) 情報処理振興事業協会技術センター：計算機用日本語基本形容詞辞書 IPAL (Basic Adjectives) —解説編, 2 技-114 (1990).
- 13) 池原 悟, 宮崎正弘, 横尾昭男：日英機械翻訳のための意味解析用の知識とその分解能, 情報処理学会論文誌, Vol.34, No.8, pp.1692-1704 (1993).
- 14) Miller, G.A., et al.: Five Papers on WordNet, CSL Report 43, Princeton University (1990, Revised 1993).
- 15) Llold, S.M.: *Roget's Thesaurus of English Words and Phrases*, Longman (1982).
- 16) 国立国語研究所：分類語彙表, 資料集 6 (1964).
- 17) 宇津呂武仁, 松本裕治：コードを用いた言語知識の獲得, 人工知能学会誌, Vol.10, No.2, pp.197-204 (1995).

(平成 7 年 7 月 27 日受付)  
(平成 8 年 10 月 1 日採録)



**横井 俊夫 (正会員)**

1941 年生. 1965 年東京大学工学部電子工学科卒業. 1966 年通商産業省工業技術院電気試験所 (現在, 電子技術総合研究所) に入所. オペレーティングシステム, 計算機アーキテクチャ, 人工知能などの研究に従事. 1982 年 (財) 新世代コンピュータ技術開発機構へ出向し第 5 世代コンピュータ・プロジェクトの推進に従事. 1987 年日本電子化辞書研究所へ出向し電子化辞書プロジェクトの推進・運営に従事. 1995 年フィリピンソフトウェア開発研修所 (PSDI) に派遣され ODA プロジェクトの推進に従事. 現在, PSDI 主席顧問. 工学博士. 電子情報通信学会, 人工知能学会, 日本ソフトウェア科学会, 日本認知科学会, 言語処理学会, 情報知識学会各会員.



**仲尾 由雄 (正会員)**

1962 年生. 1986 年東京大学理学部物理学科卒業. 同年 (株) 富士通研究所入社. 1988~1994 年日本電子化辞書研究所へ出向. 現在, (株) 富士通研究所. 自然言語処理を使った文書処理システムの研究開発に従事.



**荻野 孝野 (正会員)**

1948 年生. 1971 年東京女子大学文理学部卒業. 同年より (財) 計量計画研究所言語情報研究室勤務. 1986 年 9 月より日本電子化辞書研究所勤務, 現在に至る. 日本語情報処理の研究開発に従事. 主に, 辞書の構造化, 概念体系などの分野にかかわる. 1996 年より大東文化大学「日本文学情報処理」担当非常勤講師を兼務. 計量国語学会, 言語処理学会会員.



**田中 裕一 (正会員)**

1953 年生. 1979 年東京大学教養学部基礎科学科卒業. 同大学大学院工学系研究科情報工学専門課程中退. 1984 年富士通 (株) 入社. 1984~1992 年 (財) 新世代コンピュータ技術開発機構に出向. 1994~1995 年 (株) 日本電子化辞書研究所に出向. 1995 年 (株) ジャストシステムに入社, 現在は同社東京研究所に勤務. 専門は自然言語処理. 情報処理学会, 人工知能学会, 言語処理学会会員.