

表層レベルにおける電子化辞書の情報構造

横井俊夫^{†1,☆} 木村和広^{†2,☆}
小泉敦子^{†3,☆} 三吉秀夫^{†4,☆}

大規模な電子化辞書（その情報内容である言語知識）が表層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なく、シソーラス、コーパス、テキストベース等を含む統合的な言語情報（言語知識）のことである。表層レベルは、言語現象を分析する出発点となり、電子化辞書全体の土台となる。これに求められる情報構造は、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。そして、実現事例としてEDR 電子化辞書の表層対応部分を仕様と統計データの両面から述べる。本格的な自然言語処理技術の研究開発にとって大規模な言語データが非常に重要であることが指摘され、個別の努力が始まっている。それらの努力を統合的に把握するための電子化辞書のアーキテクチャを実現事例とあわせて提案する。

Information Structure of an Electronic Dictionary at the Surface Level

TOSHIO YOKOI,^{†1,☆} KAZUHIRO KIMURA,^{†2,☆} ATSUKO KOIZUMI^{†3,☆}
and HIDEO MIYOSHI^{†4,☆}

This paper describes the surface-level, model-information structure of a large-scale electronic dictionary that contains linguistic knowledge. The term "electronic dictionary" in this paper denotes an integrated body of linguistic information and knowledge provided by thesauri, tagged corpora, and raw corpora as well as ordinary dictionaries. The surface level is where the analysis of linguistic phenomena begins. It also constitutes the basis for the overall structure of electronic dictionaries. The required information structure was obtained by rearranging the structure of the EDR Electronic Dictionary. Next the adequacy of the information structure is proved regarding the specifications and statistical data of the EDR Dictionary at the surface level. Individual efforts have been made with a wide recognition of the importance of large-scale linguistic data for natural-language processing. The architecture of an electronic dictionary that integrates these efforts is proposed along with its instantiations.

1. はじめに

大規模な言語知識が表層レベルにおいて持つべき情報構造^{☆☆}の基本仕様を明らかにする。言語知識とは、言語に関する情報や知識のことであり、表層レベルであるとは、言語表現（言語によって表現された情報）の表記上で直接的に判断される知識に関することであ

るということである。言語知識は、電子化辞書というものに具体化される。すなわち、電子化辞書の情報内容が言語知識である。

言語知識として、語、句、文、文章、文書という言語表現の基本的な単位をすべて対象にする。すなわち、今までは、辞書（通常の辞書）であるとか、コーパスであるとか、テキストベースであるとか個別に研究開発が行われてきたものに対し、初めて統一した考え方のもとに統合的な言語知識の情報構造を明らかにするものである。これによって個々の言語知識の位置付けが明確にされ、これからの本格的な研究開発の枠組みが定まることになる。さらに、自然言語処理のための言語知識ということだけではなく、より一般的な情報

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

†2 東芝研究開発センター
Research and Development Center, Toshiba Corp.

†3 日立製作所中央研究所
Central Research Laboratory, Hitachi, Ltd.

†4 シャープ応用システム研究所
Integrated Media Laboratories, Sharp Corp.

☆ 本論文の内容は、著者等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。どう検索するのかとか、どうコンパクトに表現するのかといった実際の情報の扱いに関することは含めないようにしたものである。

や知識の構造に関する研究に対し、多くの新たな知見を与えてくれる。

情報処理技術は、今まで大規模なハードウェアや大規模なソフトウェアの構造を明らかにし、それに基づいて実用性の高いシステムを実現してきた。これからの課題として大規模な情報や知識そのものの構造を明らかにすることがある。情報や知識の入れ物としてのデータベースや知識ベースの議論から、内容（コンテンツ）としての情報や知識へと議論を進めようというものである。言語知識は、知識内容として最も一般性の高いものであり、他の知識の基底部分に位置するものである。これは、自然言語が最も汎用の知識表現言語であることに由来するものである。ただし、情報や知識の構造に関する研究はまだ初期の段階にある。ハードウェアがアーキテクチャという観点から取り組まれ始めた頃、OS等の本格ソフトの試作・研究が始まった頃に対応する。したがって、本稿では、すべての出発点という意味から、基本仕様、概念仕様というレベルに焦点を合わせる。

以下、2章では、情報構造の全体構造を説明し、表層レベルの位置付けを明らかにする。3章では、言語知識の表層レベル、すなわち電子化辞書の表層辞書の情報構造を体系的に明らかにする。その情報構造の代表的な実現事例としてEDR電子化辞書を4章に説明する。構造と計測データの両面から実証する。ただしEDR電子化辞書はあくまでも一部の第一段階としての実現である。本稿の主張はEDR電子化辞書の成果や知見を再整理し、普遍化することによって得られたものである。5章では、国内外の類似の研究開発も本稿が主張する枠組みに基本的に納まっていることを説明する。

2. 全体構造と表層レベル

言語知識の情報構造の骨格となる全体構造¹⁾を説明し、表層レベルの位置付けを明らかにする。以下では、理解のしやすさを考慮し、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一般化されたものであり、汎電子化辞書とも呼ばれる²⁾。

電子化辞書は記述の単位、記述のレベル、言語の種類³⁾の3点で、特徴付けられるサブ辞書群によって構成されている。3つの特徴を座標軸に対応させると図1のような構造体となる。

- (1) 記述のレベル：言語表現のどのレベルの知識を対象とするのかである。表層のレベルから意味記述にかかわる多様な深層レベルがある。表層レベル

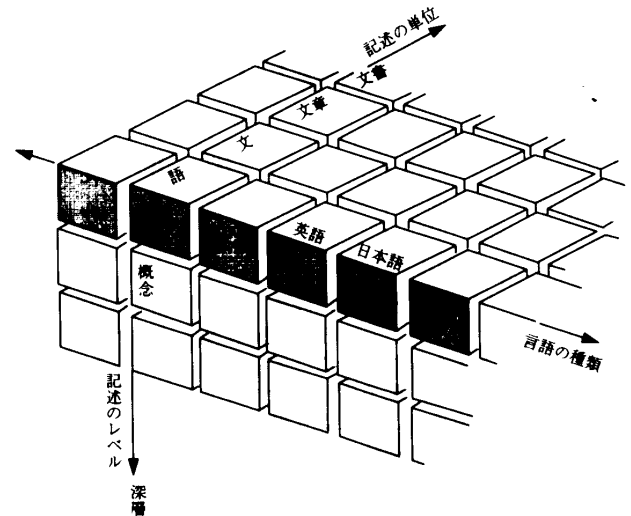


図1 全体構造と表層レベル

Fig. 1 Whole structure and surface level structure.

は、言語表現の外見的なところで判断できる知識を対象にする。形態や構文にかかわる知識は当然であるが、意味、文脈、運用にかかわるものも表記上で記述できるものは対象とする。これらは深層レベルの知識に対する判断基準ともなる。

- (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のかである。語、句、文、文章、文書という基本的な構成単位を対象とする。外見上の違いによる構成単位という考え方は、表層レベルではより効果的である。
- (3) 言語の種類：どの言語の言語表現を対象にするのかである。基本的な情報構造は言語に共通であるが、細部には言語固有のものも含まれる。

言語知識を表現するための基本単位が辞書項目である。表層レベルでは表層辞書項目である。表層レベルのサブ辞書は、表層辞書項目の集合である。表層辞書項目の基本部分はすべての構成単位に対し共通である。表層辞書項目は入れ子となる多数のサブ項目からなる。各サブ項目の情報内容には、表層辞書項目自身の情報や、他の辞書項目との関係情報が含まれる。この関係情報によって辞書項目間の関係付けが表現され、言語知識としての情報構造が形作られる。関係には、サブ辞書間にまたがる関係、辞書間関係とサブ辞書内の関係、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：記述単位の軸に沿って同じ記述レベルのサブ辞書間に定義される。上位の構成単位との要素に対応しているのかを示す。あるいは、下位の構成単位のどのような文脈となっているのかを示す。

| | |
|----------|----------------------------------|
| 〈表層辞書項目〉 | |
| 〈表記情報〉 | : 判別するための表記, 標準的なものやバリエーション等の記述 |
| 〈構成情報〉 | : 統語的役割を中心とした, 上位, 下位の構成単位との関係記述 |
| 〈要素カテゴリ〉 | : 上位の構成単位の構造の中でどのような要素となるのかの記述 |
| 〈構成構造〉 | : 下位の構成単位を要素としてどのような構造となるのかの記述 |
| 〈対応情報〉 | : 辞書間関係, 辞書内関係による他の辞書項目との対応の記述 |
| 〈意義情報〉 | : 意義の識別や基本的な定義 |
| 〈意味関係情報〉 | : 他の表層辞書項目との意味的対応 |
| 〈構成関係情報〉 | : 文脈となる表層辞書項目との対応 |
| 〈対意関係情報〉 | : 深層レベルの辞書項目との対応 |
| 〈対訳関係情報〉 | : 他言語の表層辞書項目との対応 |
| 〈環境情報〉 | : 分野, 状況等の運用に関する記述 |

図2 表層辞書項目

Fig. 2 Surface dictionary entry.

(b) 対意関係: 記述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。ある記述レベルの辞書項目からより深層の記述レベルの辞書項目で意味表現となっているものへの対応を示す。

(c) 対訳関係: 言語の種類軸に沿って同じ記述レベルのサブ辞書間に定義される。異なる言語の辞書項目間で (ほぼ) 同義であるという対応を示す。

表層レベルの言語知識が自然言語処理に果たす役割としては大きく2つある。ひとつは構文的な処理のための知識を与えることである。センテンス文法、テキスト文法、ドキュメント構造 (文法) 等の解析、生成のための体系的な情報を与える。もうひとつは翻訳や要約等の文書処理に対し、変換や対応付けのありのままではあるが安定した情報を与える。また、電子化辞書の開発プロセスの観点からは、表層レベルは既存の蓄積を利用でき、安定した言語直観を期待できるもので、言語知識の土台となる。

3. 表層辞書の情報構造

表層レベルのサブ辞書の情報構造を詳細化する。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、文書表層辞書の情報構造である。ただし、文章表層辞書、文書表層辞書については、大枠としての説明にとどめる。また、説明は日本語と英語を対象に進める。

3.1 表層辞書項目

表層辞書における情報の基本単位に対応して図2に表層辞書項目が定義される。表層辞書の情報の基本単位は次のような性質を持つものである。

- (1) 外見的に他と明確に区別しうる表記を持つこと (〈表記情報〉)。
- (2) 上位の構成構造の中でどのような要素として振る舞うのかが明確に定義できること (〈構成情報〉・〈要素カテゴリ〉)。

- (3) 対応する指示内容を確定できること (〈対応情報〉)。
- (4) 使用される環境 (状況) を規定できること (〈環境情報〉)。

通常の辞書における情報単位は、表記から検索するという前提から表記の違いのみに基づいたものである。電子化辞書においては、表記、構文、意味、運用の情報をすべて等しく見なくてはならないという観点からこのような情報単位を用いる。

3.2 語表層辞書

語表層辞書は、語を対象として表層レベルの情報を定義する。語に対応付けられる基本的な形態素としては、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。また、複数の形態素からなる派生語 (語基と派生接辞によって形成される語)、複合語 (語基と語基から形成される語)、接辞相当の固定した表現 (「において」、「に関して」等) を含め、自然言語処理の観点から語の範囲が定められる。

表層辞書項目 (図2) を語に対して詳細化した語表層辞書項目を図3に示す。〈代表表記〉^{*}は、語が実際に使用されるとき表記のバリエーションのうち、使用頻度上、あるいは語彙体系上、代表と考えられるものである。〈異表記リスト〉は、その語が実際に使用されるとき表記の揺れの一覧である。日本語では、漢字の種類、送り仮名の付け方の揺れ、漢字で表すか平仮名で表すか、外来語の場合はカタカナ表記の揺れ等、様々な異表記が生ずる。英語でも、英米の綴りの違いのほか、外来の語に多くの異表記が見られる。また、複合語に対し、ハイフンを付けるか付けないか、2語に分けるか等で表記が変わる。出現の可能性のある異表記はすべて網羅される必要がある。〈不変化部指

^{*} 通常は発音やアクセントに関する情報も含まれるが、ここでは音声に関する説明は省略する。

| | |
|-----------|--|
| 〈語表層辞書項目〉 | |
| 〈表記情報〉 | |
| 〈代表表記〉 | : 語を代表する表記. |
| 〈異表記リスト〉 | : 表記のバリエーションのリスト. |
| 〈不変部指定〉 | : 活用語に対し, その不変部分. |
| 〈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. |
| 〈構成情報〉 | |
| 〈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ときの語のクラス. |
| 〈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〉 | : 複数の形態素からなる語に対し, 語の内部構造を表示. |
| 〈対応情報〉 | |
| 〈意義情報〉 | |
| 〈語義ラベル〉 | : 語義を識別するための標識. |
| 〈語義説明〉 | : 語義の自然言語による説明文, 定義文. |
| 〈意味関係情報〉 | : 同義語, 反義語等への対応情報. |
| 〈構成関係情報〉 | : 用例文への対応情報. |
| 〈対意関係情報〉 | : 語概念への対応情報. |
| 〈対訳関係情報〉 | : 訳語への対応情報. |
| 〈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かかわる情報. |

図3 語表層辞書項目

Fig. 3 Word dictionary entry.

定〉は, 活用語の場合, その不変部, すなわち, 文字列としての不変部分を示す. 一般には語根に対応する. この情報は屈折の語形変化表として与える方法も一般的であるが, 不変部分で記述した場合は活用語尾を独立した形態素として扱うことになり, 文字列の分割・連結操作のみで形態素解析・生成を行えるという利点がある. 〈語形成〉は, その語が語基の場合に, どのような語を形成するかを示す. 主に語彙的な関係を中心に記述し, 文法的に記述可能な関係(語基と屈折接辞から形成される屈折語等)は〈接続情報〉で定義する. 「切る, 切れる, 切らす」等の動詞の自他動の対応を含む語彙的派生関係や「釣る, 釣り」等の転成も含めて扱う. これらは, 明確に語基と派生接辞に分割できなくとも, 形態的観点から見た関連語という情報を得るために記述しておく必要がある.

〈接続情報〉は, その語が文中でどのような形態素列の環境に現れるかを形態素間の接続可否の関係として表した情報である. この情報には, 自分自身のクラスを記述する方法や, 後接(あるいは前接)する形態素を直接, あるいはそのクラスによって間接的に指定する方法等がある. 自分自身のクラスのみを記述する場合には, 通常, 接続の可否を表す文法情報が, 接続テーブル等の形で必要となる. 〈統語範疇〉は, 構文的観点から見たときの語のクラスである. 広い意味での

文法範疇の一種としての品詞である. 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. しかし, 形態的観点からはすでに形態情報で接続情報として形態素の分類を行っているので, ここでは構文的観点, すなわち, 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〉は, 述語となることのできる語基に文型を記述する. 日本語では述語が要求する項(名詞句)とその格形態(格助詞)を, 英語では述語が要求する項とその位置(語順)を記述する. 本来, 文型は文の要素間関係であり, 文表層辞書に記述される情報である. したがって, 述語と項との間の意味的な選択制限, あるいは述語が項と結び付いて表現する意味は, 単語辞書の構文情報では扱わずに, 文表層辞書の〈対応情報〉で扱う. ここで記述されるのは, 述語の要求する表層的な格形態のパターンによる述語の分類である. 〈文法範疇〉は, 〈文型情報〉と同様, 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. 述語にはテンス・ムード, そのうち動詞にはさらにヴォイス・アスペクトが認められる. 日本語ではこれらの文法範疇は, 主に接辞の付加によって形態的に示される. ただし, それぞれの形態素はいくつかの異なる機能を併せ持つ場合があるため, 各形態素の接続の可否と接続する場合のカテゴリカルな意味を記述するのが

効率的である。〈語構造〉は、その語が複数の形態素からなる場合、その内部構造を表示する。

〈語義ラベル〉は、語義を識別するための標識である。たとえば、表層語に語義番号を添字したもの等である。〈語義説明〉は、自然言語による語義の説明文、定義文である。通常の辞書の語義文や用語辞典の用語説明が対応する。人間に語義の内容を正確かつ容易に想起させる役割を持つが、使用言語に適切な制限を加え、定義の明確化やコンピュータ処理を容易にする工夫をほどこすこともある。使用語彙セットを制限したり、説明文に表現構造を明示する等である。〈意味関係情報〉は、その語の類義語、反義語、上位語等を列挙することにより、語義を相対的に表現する。これだけでもある程度までの意味処理が可能となる。〈対意関係情報〉は、その語の語義と、より深い意味記述のレベルの項目、たとえば、語概念辞書項目³⁾との対応を示す。〈構成関係情報〉は、その語を含む上位の構成単位、すなわち文表層辞書項目との対応を示す。通常の辞書の用例文の列挙に対応する。〈対訳関係情報〉は、その語と、その対訳となる他言語の語(句、文)との対応関係を示す。

〈環境情報〉は、その語が使用される分野、頻度、使用者層、品格等の位相、和語、漢語等の語種等、語の運用にかかわる情報を示す。

3.3 文表層辞書

文表層辞書は、句や文を対象として表層情報を定義する。一般化、体系化されたコーパスがこれにあたる。図4に文表層辞書項目を示す。

〈文表記〉は、文を構成する文字列を定義する。意味記述の観点からは、複文や重文を文章のレベルと考

| | |
|-----------|-----------------|
| 〈文表層辞書項目〉 | |
| 〈表記情報〉 | |
| 〈文表記〉 | : 文の表記 |
| 〈構成情報〉 | |
| 〈要素カテゴリ〉 | |
| 〈文タイプ〉 | : 句、文等の統語範疇 |
| 〈構成構造〉 | |
| 〈形態素列〉 | : 文構成要素の分割 |
| 〈構文木〉 | : 文の統語構造の表示 |
| 〈対応情報〉 | |
| 〈意義情報〉 | |
| 〈文義ラベル〉 | : 文義を識別するための標識 |
| 〈文義説明〉 | : 文義の自然言語による説明文 |
| 〈意味関係情報〉 | : 同義(パラフレーズ)文等 |
| 〈構成関係情報〉 | : 文章表層辞書項目 |
| 〈対意関係情報〉 | : 文概念への対応情報 |
| 〈対訳関係情報〉 | : パラレルコーパスに対応 |
| 〈環境情報〉 | : 出典情報 |

図4 文表層辞書項目

Fig. 4 Sentence dictionary entry.

える立場があるが、ここでは、形態的な文という区別を判断基準にする。

〈文タイプ〉は、その文自身の統語範疇を定義する。句、文等の範疇名である。〈形態素列〉は、その文を構成する各形態素への分割を示す。〈構文木〉は、その文の統語構造を示す。

〈文義ラベル〉は、〈語義ラベル〉と同様に文義の識別のための標識である。〈文義ラベル〉は、文の表層表記そのものである。多義の場合は、文義を識別するための番号を付記する。〈文義説明〉は、文の意味を曖昧性のない自然言語文で説明する。たとえば、その文のパラフレーズを、可読性の高い標準的な文、すなわち、適切な語彙を用いた平易な文構造の文で定義する。〈意味関係情報〉は、同義文や反義文等の列挙である。これにより、この表層文に対しては意味解析等の重い処理を行うことなく目的の出力を得ることができる。〈構成関係情報〉は、その文が現れる文脈、つまりその文を含む上位の構成単位である文章表層辞書項目を示す。〈対意関係情報〉は、その文の文義と、より深い意味記述のレベルの項目、たとえば、文概念辞書項目との対応を示す。〈対訳関係情報〉は、他言語の訳文対応を与える。つまりパラレルコーパスの機能である。単語どうしの対応等のアラインメント情報も含まれる。

〈環境情報〉は、その文の出典情報等である。

3.4 文章表層辞書・文書表層辞書

文章および文書表層辞書は、一連の文の連鎖を対象として表層情報を定義する。ここで、文書とは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、記事等さまざまな形態をとるものであり、文章は、文書を構成する段落、段落列等を指す。いずれも、その辞書項目は、図2の基本構造に納まるものである。

文章表層辞書では、〈要素カテゴリ〉に起承転結のような文書上での役割を定義し、〈構成構造〉に文章論的修辞構造を定義する。一方、文書表層辞書では、〈要素カテゴリ〉に評論、報道記事といった文書の分野、タイプを定義し、〈構成構造〉にSGML等のマークアップ言語による文章間構造を定義する。〈対応情報〉は、文章、文書ともほぼ共通した内容を定義する。〈意義情報〉は、その文章もしくは文書をパラフレーズした文章であり、〈意味関係情報〉は、要約文章(文書)等であり、〈対意関係情報〉は、文章(文書)概念辞書項目への対応情報であり、〈対訳関係情報〉は、アラインメント情報を含む他言語の翻訳文章(文書)への対応である。文章表層辞書では、〈構成関係情報〉にその文章の出典となる文書表層辞書項目が示される。ただし、文書表層辞書ではそれ以上の記述の単位を仮定せず、

| | |
|---------------|-------------------------|
| 〈日本語単語辞書レコード〉 | |
| 〈レコード番号〉 | :レコードタイプと識別番号 |
| 〈見出し情報〉 | |
| 〈単語見出し〉 | :活用語尾つき見出し表記とその読み |
| 〈不変化部-連接属性対〉 | :構成語の不変化部とその連接属性の対の列 |
| 〈かな表記〉 | :不変化部のカタカナによる表記 |
| 〈発音〉 | :不変化部のカタカナ表記による発音とアクセント |
| 〈文法情報〉 | |
| 〈品詞〉 | :品詞 |
| 〈構文本〉 | :構成語の係受け関係(慣用句のみ) |
| 〈活用情報〉 | |
| 〈活用形情報〉 | :活用形(不規則活用語のみ) |
| 〈活用型情報〉 | :活用型(規則活用語のみ) |
| 〈表層格情報〉 | :表層格(用言, 述語句のみ) |
| 〈相情報〉 | :相情報(動詞のみ) |
| 〈機能語情報〉 | :機能語情報(機能語のみ) |
| 〈意味情報〉 | |
| 〈概念識別子〉 | :概念の同一性を示す番号 |
| 〈概念見出し〉 | :概念を代表する単語見出し(内容語のみ) |
| 〈日本語概念見出し〉 | :概念を代表する日本語単語見出し |
| 〈英語概念見出し〉 | :概念を代表する英語単語見出し |
| 〈概念説明〉 | :概念の文章による説明(内容語のみ) |
| 〈日本語概念説明〉 | :概念の日本語による説明 |
| 〈英語概念説明〉 | :概念の英語による説明 |
| 〈運用・その他情報〉 | |
| 〈用法〉 | :語の用法 |
| 〈頻度〉 | :出現頻度 |
| 〈管理情報〉 | |
| 〈管理履歴レコード〉 | :更新日付等の管理情報 |
| 〈日英対訳辞書レコード〉 | |
| 〈対訳情報〉 | |
| 〈訳語種別〉 | :訳語の種類 |
| 〈訳語表記〉 | :訳語の表記 |
| 〈訳語品詞〉 | :訳語の品詞 |

図5 EDR 電子化辞書の語表層辞書対応部分

Fig. 5 Records of the EDR Electronic Dictionary corresponding to the word dictionary.

定義しない。

4. EDR 電子化辞書における実現

表層辞書の情報構造の実現事例としてEDR 電子化辞書⁴⁾の表層辞書対応部分を説明する。EDRの表層辞書は、既存の機械可読辞書(MRD)とは異なり、言語処理上の要請にのみ基づいて設計・開発されたものである。したがって、これを検証することで、情報構造の妥当性を立証するとともに、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様を説明する。

4.1 辞書仕様

EDR 電子化辞書では、単語辞書、対訳辞書、共起辞書、コーパス、テキストベースの5種の辞書によって表層辞書を実現する。言語の種類は、日本語と英語の2種である。

語表層辞書を基本的に実現したものが単語辞書と対訳辞書である。2つの辞書への分離は現実的要請に

よる。単語辞書は、語表層辞書の主要部分を担い、その仕様をほぼ実現している。対訳辞書は、文字どおり語表層辞書の〈対訳関係情報〉を与える。対訳辞書では、訳語のほかに源言語との意味的關係を示す訳語種別等の情報も定義されている。この有様を、日本語単語辞書の辞書レコードの仕様と日英対訳辞書の辞書レコードの仕様の一部を図5に示すことによって表す。EDR 電子化辞書では、辞書項目に対応するものを辞書レコードと呼ぶ。

表層辞書項目における〈表記情報〉、〈構成情報〉、〈対応情報〉、〈環境情報〉は、EDR 辞書レコードでは、それぞれ、〈見出し情報〉、〈文法情報〉、〈意味情報〉、〈運用・その他情報〉に対応している。このうち、〈対応情報〉については〈対意関係情報〉に対応する部分が〈意味情報〉として、〈対訳関係情報〉に対応する部分が〈対訳情報〉として各々実現されている。〈意味情報〉は、概念レベルの語辞書であるEDR 概念辞書と共有している情報であり、単語表層辞書項目と単語概念辞書項

| | |
|---------------|------------------------------|
| 〈日本語コーパスレコード〉 | |
| 〈レコード番号〉 | :レコードタイプと識別番号 |
| 〈文情報〉 | |
| : 〈テキスト番号〉 | :文管理番号 |
| 〈出典情報〉 | :出典テキスト名 |
| 〈文〉 | :用例文の表記 |
| 〈構成要素情報〉 | |
| 〈構成要素列〉 | :形態素, 構文, 意味情報が共有する文構成要素のラベル |
| 〈形態素情報〉 | |
| 〈形態素列〉 | :文の形態素分割 |
| 〈構文情報〉 | |
| 〈構文木〉 | :文の構文構造 |
| 〈管理情報〉 | |
| 〈管理履歴レコード〉 | :更新日付等の管理情報 |

図6 EDR 電子化辞書の文表層辞書対応部分

Fig. 6 Records of the EDR Electronic Dictionary corresponding to the sentence dictionary.

目との直接の対応付けを行うことにより、辞書間の連続性を保証するものである。一方、〈意義情報〉、すなわち語義は、辞書開発の初期段階で〈意味情報〉と同じ形式で定義されたものである。語に対して定義された語義は、概念レベルの情報記述が目的とする分解能にあわせて、同義のものが1つに統合され1つの概念に近似される。〈意味情報〉はこの近似の結果である。なお、〈意義情報〉は辞書仕様上未公開としている。これは、〈意味情報〉との区別が一般利用者にとって困難であることへの配慮と、〈意味情報〉のだけでも実用上差し支えないとの判断による。単語辞書および対訳辞書は、語表層辞書項目のうち、当面必要なものはすべて実現されているが、〈代表表記〉と〈異表記リスト〉の区別、〈語形成〉、〈意味関係情報〉、〈構成関係情報〉の拡充については今後委ねられている。

文表層辞書の基本部分の実現となるのが、共起辞書とコーパスである。ただし、共起辞書は、現実的要請から、コーパスの情報の一部である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。EDRコーパスの辞書レコードを図6に示す。文表層辞書項目における〈文表記〉はEDRコーパスでは〈文〉で、〈構成構造〉は〈構成要素情報×形態素情報×構文情報〉で、〈環境情報〉は〈出典情報〉で、それぞれ実現されている。〈構文情報〉は依存文法に基づいたものである。〈対応情報〉は基本的に実現の対象にはしていない。ただし、〈構成関係情報〉については〈テキスト番号〉というEDRテキストベースへのリンクで暗黙に表現されている。なお、EDRコーパスにおける〈意味情報〉は、概念レベルの情報であり本稿の〈対応情報〉の範囲ではない。EDRコーパスは、単言語コーパスである。今後は〈対訳関係情報〉の付加も望まれる。また、もともとコーパス開発は一義的目標でなかったため、収集したテキストの出典に偏りがある

等の課題が残されている。

文章・文書表層辞書に対応するものとしてテキストベースがある。ただし、現在実現されているのは〈表記情報〉と簡単な〈構成情報〉のみである。

4.2 統計データ

語表層辞書に相当するEDR電子化辞書の統計データのうち、単語辞書および対訳辞書に関するものを表1と表2に示す。表1は日本語基本語、表2は英語基本語にかかわるものである。ただし、仕様上対象としていない項目（機能語等実体概念を持たない語に対する平均概念数等）は、“—”で示す。

表の構成比とは、単語数等が全体数に占める割合を百分率で示したものである。日本語基本語に関しては、一般の冊子対辞書の見出し語の品詞分布とは若干の違いが見られ、名詞に比べ動詞や形容動詞の比率が高い。一般に大規模の冊子体辞書では、一般語のほかに百科語や固有名詞等が多く含まれるため、名詞の構成比が80%を超えることも多い。EDR電子化辞書では固有名詞の登録を最小限にとどめていることと、品詞認定に対する姿勢の違いが主たる要因である。固有名詞は多種多数になるが、一般的なものは拡充は容易であり個別のものは利用者に任せるべきであるとの判断からである。品詞認定に関しては、通常名詞としてのみ分類される語も、用言化（サ変動詞化）して用いられるものについては用言としての分類も行っている。これは用言としての振舞いも記述しておく必要があるからである。このぶん用言の構成比率が高くなっている。平均語義数、平均概念数とは、各々単語あたりの語義数、概念数であり、意味分割の細かさ、粒度を表す指標となる。平均語義数と平均概念数との比は、語義として意味分割したものが語概念としてどの程度同一の概念として近似されているかを示す。全体的に英語より日本語の方が統合（同一化）が進んでいることが読

表1 EDR 日本語単語辞書・日英対訳辞書諸元
Table 1 Statistics for EDR Japanese Word Dictionary and Japanese-English Bilingual Dictionary.

| 品詞類 | 単語数 | 構成比 (%) | 平均語義数 | 平均概念数 | 語義あたり訳語数 | 単語あたり訳語数 |
|-------|---------|---------|-------|-------|----------|----------|
| 名詞類 | 189,042 | 73.2 | 1.46 | 0.81 | 1.96 | 1.06 |
| 動詞類 | 44,556 | 17.3 | 1.73 | 0.69 | 3.09 | 1.50 |
| 形容詞類 | 2,161 | 0.8 | 1.73 | 0.73 | 3.70 | 1.59 |
| 形容動詞類 | 7,937 | 3.1 | 1.52 | 0.65 | 4.61 | 1.46 |
| 副詞類 | 4,634 | 1.8 | 1.60 | 0.65 | 3.46 | 1.76 |
| 連体詞類 | 368 | 0.1 | 1.49 | 0.74 | 3.70 | 1.46 |
| 接続詞類 | 378 | 0.2 | 1.15 | - | - | - |
| 接頭語類 | 349 | 0.1 | 1.65 | 1.16 | 3.18 | 2.47 |
| 接尾語類 | 1,719 | 0.7 | 1.38 | 0.87 | 2.18 | 1.27 |
| 語尾類 | 91 | 0.0 | 1.36 | - | - | - |
| 構文要素類 | 5,783 | 2.2 | 1.26 | 0.89 | - | - |
| 助詞類 | 189 | 0.1 | 1.01 | - | - | - |
| 助動詞類 | 228 | 0.1 | 1.07 | - | - | - |
| 補助用言類 | 252 | 0.1 | 1.50 | 0.59 | 2.64 | 1.44 |
| 感動詞類 | 574 | 0.2 | 1.44 | 0.77 | 2.50 | 1.23 |
| その他 | 61 | 0.0 | 1.00 | 0.80 | 4.26 | 3.84 |
| 総計 | 258,322 | 100.0 | 1.51 | 0.78 | 2.27 | 1.03 |

表2 EDR 英語単語辞書・英日対訳辞書諸元
Table 2 Statistics for EDR English Word Dictionary and English-Japanese Bilingual Dictionary.

| 品詞類 | 単語数 | 構成比 (%) | 平均語義数 | 平均概念数 | 語義あたり訳語数 | 単語あたり訳語数 |
|-------|---------|---------|-------|-------|----------|----------|
| 名詞類 | 116,896 | 61.3 | 1.56 | 1.23 | 1.35 | 1.25 |
| 代名詞類 | 174 | 0.1 | 1.37 | 0.59 | 1.05 | 0.96 |
| 動詞類 | 14,573 | 7.6 | 3.29 | 1.97 | 1.70 | 2.77 |
| 形容詞類 | 38,168 | 20.0 | 1.52 | 1.30 | 1.52 | 1.36 |
| 副詞類 | 7,846 | 4.1 | 1.28 | 1.10 | 1.61 | 1.13 |
| 前置詞類 | 221 | 0.1 | 1.00 | - | - | - |
| 決定詞類 | 30 | 0.0 | 1.10 | - | - | - |
| 助動詞類 | 96 | 0.1 | 1.32 | - | - | - |
| 間投詞類 | 380 | 0.2 | 1.19 | 0.86 | 1.62 | 0.91 |
| 接続詞類 | 19 | 0.0 | 1.00 | - | - | - |
| 語尾類 | 173 | 0.1 | 1.00 | - | - | - |
| 構文要素類 | 12,148 | 6.4 | 1.19 | 0.39 | 0.78 | 0.35 |
| その他 | 118 | 0.1 | 1.00 | 0.41 | 0.68 | 0.68 |
| 総計 | 190,842 | 100.0 | 1.65 | 1.24 | 1.42 | 1.30 |

み取れる。英語の語義に対しては同義性の判断が難しく近似の度合に差が生じたものと思われる。

語義あたり訳語数は、訳語の延べ総数を語義数で割った値であり、単語の語義に対して平均いくつの訳語が与えられているのかを示す。単語あたり訳語数は、訳語の異なり数を単語数で割った値であり、訳語を、対訳言語を意味表現言語とした意味記述と見ると、意味分割の粒度を表す指標となる。語義あたり訳語数を日本語と英語で比較してみると、全般に日英対訳の方が訳語数が多い。これは訳語を記述していくうえでの言語による違いを反映していると考えられる。日本語は造語能力が高いので、英日対訳における日本語による訳語は複合語表現によって適切な一語で表すことが比較的容易である。一方、日英対訳では英語で同義性を満たすような訳語を記述しようとすると複雑な句や

文で表現せざるを得なくなり、これを避けるために類義語の列挙による記述が多くなる。このことは、単語あたりの訳語数で比較したときに英日対訳の方が訳語の数が多いことにも反映されている。

平均語義数、単語あたり訳語数、平均概念数を比較してみると、日本語、英語ともにこの順番で小さくなる。これは、訳語の異なりを意味の粒度の基準として見たときに、語義はより細かい粒度で、概念はより粗い粒度で記述されていることを表している。これはEDR電子化辞書の1つの特徴といえる。

文表層辞書に相当するEDR電子化辞書の統計データとして、EDRコーパスに関するものを日本語について表3、英語について表4に示す。表の品詞は形態素に対するものである。複合語に対してはそれを構成する形態素の品詞を集計してある。出典別に見た場合

表3 EDR 日本語コーパス諸元
Table 3 Statistics for EDR Japanese Corpus.

| 品詞類 | 形態素数 (総数) | 構成比 (%) | 形態素数 (異なり数) | 構成比 (%) |
|-------|-----------|---------|-------------|---------|
| 名詞類 | 1,375,378 | 26.1 | 110,912 | 81.3 |
| 動詞類 | 622,125 | 11.8 | 14,638 | 10.7 |
| 形容詞類 | 58,742 | 1.1 | 1,204 | 0.9 |
| 形容動詞類 | 61,192 | 1.2 | 3,796 | 2.8 |
| 副詞類 | 74,332 | 1.4 | 2,934 | 2.1 |
| 連体詞類 | 40,271 | 0.8 | 247 | 0.2 |
| 接続詞類 | 23,562 | 0.4 | 247 | 0.2 |
| 接頭語類 | 21,063 | 0.4 | 318 | 0.2 |
| 接尾類 | 122,954 | 2.3 | 1,330 | 1.0 |
| 語尾類 | 631,304 | 12.0 | 155 | 0.1 |
| 助詞類 | 1,402,757 | 26.7 | 171 | 0.1 |
| 助動詞類 | 319,852 | 6.1 | 203 | 0.1 |
| 感動詞類 | 356 | 0.0 | 105 | 0.1 |
| その他 | 508,333 | 9.7 | 226 | 0.2 |
| 総計 | 5,262,221 | 100.0 | 136,486 | 100.0 |

表4 EDR 英語コーパス諸元
Table 4 Statistics for EDR English Corpus.

| 品詞類 | 形態素数 (総数) | 構成比 (%) | 形態素数 (異なり数) | 構成比 (%) |
|------|-----------|---------|-------------|---------|
| 名詞類 | 660,027 | 23.6 | 59,091 | 72.8 |
| 代名詞類 | 185,208 | 6.6 | 391 | 0.5 |
| 動詞類 | 369,583 | 13.2 | 8,338 | 10.3 |
| 形容詞類 | 188,115 | 6.7 | 9,217 | 11.4 |
| 副詞類 | 130,193 | 4.7 | 2,251 | 2.8 |
| 前置詞類 | 289,064 | 10.3 | 273 | 0.3 |
| 決定詞類 | 204,199 | 7.3 | 36 | 0.0 |
| 助動詞類 | 65,213 | 2.3 | 109 | 0.1 |
| 間投詞類 | 1,856 | 0.1 | 158 | 0.2 |
| 接続詞類 | 88,709 | 3.2 | 148 | 0.2 |
| 接辞類 | 198 | 0.0 | 53 | 0.1 |
| 語尾類 | 336,489 | 12.0 | 192 | 0.2 |
| その他 | 274,759 | 9.8 | 939 | 1.2 |
| 総計 | 2,793,613 | 100.0 | 81,196 | 100.0 |

の構成は、日本語では、新聞 43%、雑誌 32%、教科書 14%、事典 11%であり、英語では、新聞 65%、用例集 35%である。形態素総数で見た品詞の構成比は書き言葉の文構造を反映した比率となっている。特に、日本語では、話し言葉に比較すると副詞、感動詞の比率が小さい。形態素の異なり数で見た場合は、日本語、英語ともに名詞の比率が非常に大きく、オープンクラスの単語であることが反映されている。

その他の EDR 電子化辞書の相当部分については、実現規模のみを付記しておく。共起辞書の規模は、日本語で 115 万タプル、英語で 60 万タプルである。テキストベースの規模は、日本語で 2000 万文、英語で 500 万文である。

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、

表層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コーパスとして分類されるものがある。

辞書として分類される事例として、欧州の Acquilex, Multilex, Genelex, 米国の Comlex, 日本の IPAL 辞書, ALT 辞書, JICST 対訳辞書を取りあげる。

Acquilex⁵⁾は、機械可読辞書 (MRD) からの半自動的な語彙知識の獲得を目的としている。最終的に獲得される語彙知識項目は、統語情報と意味情報が同じ枠組み (タイプつき素性構造) で記述され、単一化ベースの自然言語処理で直接利用できるものである。語彙知識項目は、MRD (正確にはその内容を分析し構造化したデータベース) の持つ情報をあらかじめ定義したタイプシステムを通じて展開することにより得られる。主な関心は、既存の MRD に内包された知識をいかに形式化し利用するかであり、本稿や他の事例のように、MRD が内包しない知識をも対象として電

子化辞書に要請される情報構造を検討し、設計開発を行っていく立場とは異なる。

Multilex⁶⁾は、多言語間の辞書データベースの基準を設け、実験することを目的としている。Multilexデータベースは、言語ごとに1つの単語辞書と言語対ごとに2つの対訳辞書から構成される辞書の集合体である。Multilexで提案されている単語辞書は、Lexical Units (LU) と呼ばれる単位で構成される。各LUは、表記・音韻・形態の単位 (GPMU)、統語情報、意味情報、相互参照 (同義語、反義語等へのリンク情報)、変換 (対訳辞書項目へのリンク情報)、管理レコードを持ち、本稿で提案した語表層辞書の情報構造に最も忠実な構造を持っている。

Genelex⁷⁾は、共通の辞書仕様のもとで、様々なヨーロッパ言語に対する汎用辞書を開発することを目的としている。特定の言語処理応用システムは想定しない。辞書項目の単位は、形態単位と呼ばれ、表記が同じでも品詞が異なれば分割される。各形態単位は、1つ以上の統語単位と意味へのリンクを持つ。汎用性を重視しあまり複雑な情報構造は持たないが、〈表記情報〉と〈構成情報〉の主要部分は充実しており、また大規模な実現を見ている。

Comlex⁸⁾は、大規模テキスト処理研究のための共有資源となることを目的とし、商用を含めて広く一般に提供されているものである。既存のMRDを使ってひな型を作り、その後人手作業を加える戦略により、開発効率が非常に高い。Comlexの特徴は、意味の領域には踏み込まず、ユーザに共通の理解がある構文構造の記述、特に、〈文型情報〉にあたる下位範疇化素性の定義に的を絞っていることである。ただし、意味情報については、WordNet⁹⁾との統合を狙っている。また、収録対象は、オープンクラスの単語、すなわち、名詞、動詞、形容詞、副詞に焦点をあて、他の品詞についてはあまり重要視していないことも特徴である。word structure と呼ばれる辞書項目は、見出し表記、品詞、語形、テンス、単複、下位範疇化素性、文法属性、変更履歴からなり、本稿で述べた情報構造の部分的実現にとどまっただけのもの、すばやい提供を可能としたことにその意義がある。

IPAL 辞書¹⁰⁾は、語数を制限し基本的な語についてのみではあるが、非常に綿密な分析と記述を行っている。特に、〈文型情報〉と〈文法範疇〉の実現は充実している。

ALT 辞書 (機械翻訳システム ALT 用辞書)¹¹⁾は、自然言語処理、特に機械翻訳において、表層に近いレベルの意味解析を実用規模で実現することを目的とし

て開発されている。ALT 辞書では、表現の仕方 (表層構造) を意味の一部ととらえており、これを端的に表す構文意味辞書と呼ばれる用言の辞書は、語に対して定義されるものの、慣用文型を含む豊富な文型パターンに対し、文の意味、本稿でいう文義を分節している点で、思想的には文表層辞書と文概念辞書を一体化したものを部分的に実現していると考えられる。

JICST 対訳辞書¹²⁾は、日本語科学技術論文の抄録を翻訳するために利用されている。目的を特化したことにより、必要な記述項目を限定し、語表層辞書の実現を行った事例といえる。

コーパスとして分類される事例として、米国の Penn Treebank と英国ナショナルコーパスを取りあげる。

Penn Treebank¹³⁾は、書き言葉、話し言葉の両方を含み、系統的に収集されたブラウンコーパス以外に、新聞、官公庁資料、マニュアル等、様々な文章を収集している。付加されたタグは、品詞と構文構造が中心であり、文表層辞書の部分的実現事例といえる。

英国ナショナルコーパス¹⁴⁾は、辞書の編纂、言語学研究、言語処理等、幅広い利用を目的として収集されている。コーパス内のテキストでは、表題、段落、脚注等を表すのに SGML を使用する。またソフトウェアで自動生成されるタグも SGML で表す。また、テキストの出典にかかわる書誌的事項には、テキスト執筆者の性、年齢層等まで記載されており、充実した内容となっている。文章・文書表層辞書の主要部分を実現した事例である。

6. ま と め

統合的な言語データ (表層レベルの言語情報) の情報構造によって個々の言語データの役割や相互の関係が明らかになる。これによって、実際の言語データ開発のそれぞれの努力が歩調を合わせていく共通の基盤が生まれることになる。

自然言語処理からの言語データへの要求は、語から文、文章へと拡大しつつあり、語や文に対しては、網羅性の達成や精度向上に向かいつつある。処理の単位を大きくするに従って、一般的には処理の機能や精度の向上が達成される。もちろん、このために通常は処理の負荷が増し、適切な妥協点を求めることによって実用化がはかられる。実際の言語データの開発は、自然言語処理技術や応用システムの研究開発、さらには商用化等の活動と適切なバランスを保ちつつ進められることになる。本稿で述べた電子化辞書の情報構造は、今後の自然言語処理技術の基本的な技術動向をふまえて設計されたものである。したがって、これからの言

語データの開発・整備の努力は自ずからこの枠組みにまとめ上げられていくものと期待される。

謝辞 EDR 電子化辞書プロジェクトに携わった方々に感謝する。

参考文献

- 1) Yokoi, T. and Yasuhara, H.: The Information Structure of Linguistic Knowledge, *Information Systems* (to be submitted)
- 2) 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, 第1回言語処理学会年次大会論文集, pp.185-188 (1995)
- 3) 横井俊夫, 仲尾由雄, 荻野孝野, 田中裕一: 概念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 (投稿中)
- 4) 日本電子化辞書研究所: EDR 電子化辞書仕様説明書 (第2版), EDR TR-045 (1995) [<http://www.ijnet.or.jp/edr> で ftp 可能]
- 5) Briscoe, T., et al.: ACQUILEX: Acquisition of Lexical Knowledge for Natural Language Processing Systems, Esprit BRA-3030 Periodic Progress Report No.1 (1990)
- 6) Serasset, G.: Recent Trends of Electronic Dictionary Research and Development in Europe, EDR TM-038 (1994)
- 7) Normier, B. and Nossin, M.: GENELEX Project: EUREKA for Linguistic Engineering, *Proc. Int'l Workshop on Electronic Dictionaries* (EDR TR-031) (1991)
- 8) Grishman, R., Macleod, C. and Meyers, A.: Complex Syntax: Building a Computational Lexicon, *Proc. 15th Int'l Conf. Computational Linguistics* (COLING94), pp.268-272 (1994)
- 9) Miller, G., et al.: Five Papers on WordNet, CSL Report 43, Princeton University (1990, Revised 1993)
- 10) 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) — 解説編一, 61 技-073 (1987)
- 11) 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)
- 12) 芦崎達雄: JICST 機械翻訳システム (1), 機械翻訳辞書の作成, 情報管理, Vol.37, No.1, pp.27-43 (1994)
- 13) Marcus, M., Santorini, B. and Marcinkiewicz, M.: Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, Vol.19, No.2, pp.313-330 (1993)
- 14) BNC: British National Corpus - Written Corpus Design Specification (1991)

(平成7年7月31日受付)

(平成7年12月8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, 人工知能等の研究に従事。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(株)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。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, 言語処理学会各会員。



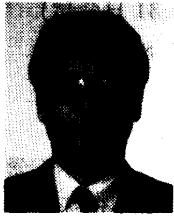
木村 和広 (正会員)

1959年生。1981年早稲田大学理工学部電子通信学科卒業。同年東京芝浦電気(株)入社。1985~1988年(財)新世代コンピュータ技術開発機構出向。1992~1995年(株)日本電子化辞書研究所出向。現在(株)東芝研究開発センター情報・通信システム研究所。自然言語処理システム, 電子化辞書の研究開発に従事。ACL, 言語処理学会各会員。



小泉 敦子 (正会員)

1958年生。1982年国際基督教大学教養学部語学科卒業。同年(株)日立製作所入社。現在, 同社中央研究所に勤務。入社以来, 機械翻訳システムの研究開発に従事。1992~1995年(株)日本電子化辞書研究所出向。



三吉 秀夫 (正会員)

1954年生. 1977年大阪大学基礎工学部情報工学科卒業. 1979年同大学大学院基礎工学研究科修士課程修了. 同年シャープ(株)入社. 1982~1987年(財)新世代コンピュータ技術開発機構出向. 1989~1990年Stanford大学CSLIに留学. 1992~1995年(株)日本電子化辞書研究所出向. 現在シャープ(株)マルチメディア開発本部応用システム研究所. 自然言語の意味処理, 対訳辞書の研究開発に従事. 人工知能学会, 日本認知科学会各会員.
