単語意味辞書と単語体系

横尾 昭男[*1] 宮崎 正弘[*2] 白井 諭[*3] 阿部さつき[*4]
小船 園望[*4] 池原 悟[*5] 大山 芳史[*3] 小倉健太郎[*3]

[*1]ATR 音声翻訳通信研究所 [*2]新潟大学 工学部 [*3]NTT コミュニケーション科学研究所
[*4]NTT アドバンステクノロジ [*4]鳥取大学 工学部

1 はじめに

日英機械翻訳の意味解析の精度向上を目指し，日本語の新聞等の実用文で使用される 40 万語に意味属性を付与した単語意味辞書の構築を進めてきた[池原 93]。この意味属性は，単語の意味的な用法に着目することにより，一般名詞が 2,700 種類，固有名詞が 130 種類に分類・体系化されている。先に上梓した単語体系[池原 97]は，単語の意味を詳細に分類するため，単語意味辞書の主として意味属性に関する部分を，人間が利用することを意識して編集し直したものである。本稿では，まず単語意味辞書から単語体系への編集の概要を報告する。次に，この編集を通して，特に意味属性別単語表の作成を通して明らかになった単語意味辞書の課題を整理し，今後の取り組みについて述べる。

2 ALT-J/E におけるシステム辞書の構成

ALT-J/E のシステム辞書は図 1 のように構成されている。

意味属性体系は，対象の見方や捉え方，すなわち，対象を概念化する際の視点を整理したものであり，話者から見れば単語の用法に相当する分類となっている。具体的には，分類観点として，上位－下位(is-a)関係のほかに，全体－部分(has-a)関係にも着目し，階層的な木構造にまとめたものである[池原 97,宮崎 97,中岩 97]。一般名詞意味属性体系は，固有名詞を除くすべての名詞の体系的な分類を行なうためのものである。固有名詞意味属性は，人名，地名等の詳細かつ体系的な分類を行なうためのものであり，一般名詞意味属性の一部を詳細化したような構成となっている。用言意味属性は，用言そのものというよりも格要素を含めた文型の体系的な分類を行なうためのものである。これらの意味属性体系は，日英翻訳における訳し分けや文型記述を通して妥当性を検証してきた。

以下では，単語意味辞書について概要を述べる。単語意味辞書の記述には，一般名詞意味属性体系と，固有名詞意味属性体系が使用される。

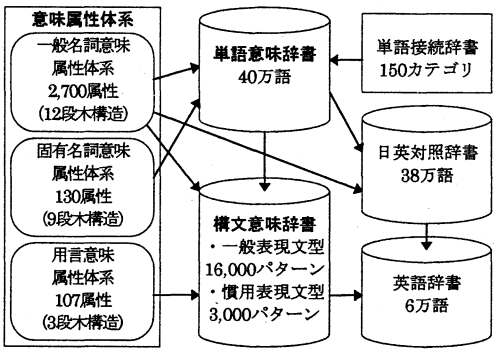

図 1 ALT-J/E におけるシステム辞書の構成

3 単語意味辞書の概要[池原 97, 横尾 97]

3.1 見出し語の収録条件

(1)収録対象語

実用的な日英機械翻訳を実現するため，現代国語の記述文で使用される単語を網羅的に収録することを目指した。国語辞書に収録されている一般語のほか，新聞記事で多用される人名，地名，企業名などの固有名詞や，技術文書で使用される専門用語を収録対象とした。

(2)単語単位

日本語は造語力が高く，複合語が際限なく生成される。これらを網羅的に収録するのは不可能である。そこで，原則として，単語辞書には，基本語や接辞などの短単位語（語基）を収録することとした。派生語や複合語などの長単位語は，解析処理により短単位語の組み合わせに分割して扱う[宮崎 84,宮崎

93]。ただし，適切な分割が困難な複合語や，「冷暖房」のような縮退語は，必要に応じて例外的に長単位で収録した。

(3)表記の揺れ

日本語には明確な正書法が確立されておらず，同じ語を漢字，ひらがな，カタカナ，およびそれらの混ぜ書きなどで表記したり，送りがなが揺らぐといった問題がある。これらの揺らぎを処理で対処するのは容易でなく，また，すべてを辞書に登録すると単語情報の記述が膨大かつ冗長なものになる。そこで，揺らぎが考えられる場合は，それらのうちの代表的な表記を1つ選定し，単語意味辞書にはあらゆる変化形を登録するが，それ以外の辞書には代表形のみを登録することにした。記述内容の整合性を保つため，辞書メンテナンスツールにより単語情報の共通化を支援している。

(4)活用語

規則的に活用する語は不変化部分を登録し，不規則に活用する語はすべての変化形を登録する。ただし，形態素解析における1文字単語候補を抑制するため，不変化部分が1文字である動詞はすべての変化形を登録した。また，助動詞は，規則的に活用する語も含め，すべての変化形を登録した。

3.2 見出し語の収集経過

(1)一般語

人間用の国語辞書は，人間が持つ常識や類推能力を前提として編集されている。このため，機械用の辞書を編集するにあたっては，国語辞書から，見出し語だけでなく，子見出しや派生語を収集対象に加えたほか，語彙調査資料（国立国語研究所『電子計算機による新聞の語彙調査』），人間用のシソーラス（国立国語研究所『分類語彙表』，『角川類語新辞典』），新聞記事（日本経済新聞）1年分を対象とした独自の語彙調査からも語の補充を図った。収集語数は12万語である。

(2)固有名詞

地名としては，いわゆる地名のほか，日本の行政区画名，駅名，国際地域名，国名，都市名などを収集した。『理科年表』等から天体名を収集した。

人名としては，日本人の姓，日本人の名，歴史上の有名人名，現代の有名人名，神仏名を収集した。

組織名としては，『会社四季報』等から日本の企業名を収集したほか，日本の官庁名，団体・党派名，日本の大学・短大・高専，国際機関名，条約名を収集した。

収集規模は，異なりで20万語である。

(3)専門用語

機械翻訳の実験計画を考慮して，電気電子情報関連の専門用語を主たる収集対象とした。具体的には，日英対訳の『科学技術用語25万語辞典』から5万語を収集した。また，新聞に頻出する時事用語をカバーするため，『現代用語の基礎知識』，『imidas』，『記者ハンドブック』等から3万語を収集した。

3.3 単語の記述情報

単語ごとに，品詞や単語連接などの文法情報のほか，意味属性や共起情報などの意味や用法に関する情報を記述する。収録語数の増加に伴って必要性が顕在化してきた同形の語や表現を識別するための情報も記述する。主な記述情報を表1に示す。

表1 単語に対する主な記述情報

表記	見出し	見出しの字面
	標準表記	見出し語の表記上の代表形
	読み	見出し語の発音のひらがな表記
音韻情報		見出し語の発音の連濁化の可否
形態情報		漢字，ひらがな，カタカナ，英数字，混ぜ書きの別
文法情報	品詞補足情報	自動詞／他動詞の別
		本動詞／補助動詞の別
		意思動詞／無意思動詞の別
		形容動詞型名詞派生形（「～な／～に」の可否）
		副詞派生形（「～と／～に／～で」の可否）
		引用の「と／か」の可否
	前方（後方）接続カテゴリ	見出し語の直前（直後）に文法的に接続可能な品詞のグループを示すコード
	時制情報	瞬間／継続／情態などの別
	様相情報	副詞・助動詞等の様相属性
	接続情報	接続詞・助動詞等の接続属性
意味属性情報		一般名詞意味属性（固有名詞にも付与）
		固有名詞意味属性（固有名詞のみ）
位相	語種情報	一般語／固有語／専門語の別
	使用分野	専門分野種別，文書形式種別
重要度	重要度	最重要語／重要語／基礎語の別
	統計的情報	頻度統計結果による見出し語の出現頻度
単語優先属性		同形語の選択方法と優先度
辞書検索制御	最長単語情報	見出し語を包含する語がないことを表示
	同形最終情報	一連の同形語の最終レコードに付与
	見出し語内単語連鎖情報	見出し語に包含されている単語があることを表示

4 単語体系の編集

4.1 編集方針

単語意味辞書に収録されている情報は，日英機械翻訳における日本語解析に必要となるものである。日本語におけるいわば「常識」を分類定量化していると言える。このため，人間用の国語辞書にはなじみのない情報が記述されたり，なじみのある情報でも，細かく分類されていたり，なじみのない形式で記述されたりしている。

これまでの国語辞書には記述されなかった情報が多く含まれているので，適切な形式で編集すれば，新しいタイプの辞書として利用されることが期待できる。一方，単語意味辞書に収録されている情報は多岐に渡るため，これらを一律に編集しても，辞書の目的がぼやけてしまう恐れがある。

そこで，国語辞書は主に単語の意味を調べるために使用される点に鑑みて，意味属性情報を中心に編集する方針を立てた。すなわち，単語体系では，現代国語の記述文で使用される単語を収録対象とし，単語ごとにどのような意味属性が付与されているかを一覧形式で表示することにした。

具体的には次のようにした。単語の配列は，国語辞書と同様に単語の読みに基づくものとし，単語に異表記や表記の揺らぎがある場合は原則としてすべてを併記する。単語の種別を示すため，学校文法の 10 品詞を若干詳細化した 21 品詞を付与する。そして，名詞系の単語に意味属性を付与するが，固有名詞意味属性は一般名詞意味属性を詳細化したものであるので，これらが重複するのを避け，固有名詞には固有名詞意味属性を，それ以外の名詞には一般名詞意味属性を付与する。

表 2　単語体系の品詞一覧

名詞系	一般名詞，サ変名詞，形容動詞型名詞，転生名詞，時詞，数詞，形式名詞，固有名詞，代名詞，接頭語，接尾語
その他	自動詞，他動詞，補助動詞，形容詞，連体詞，副詞，接続詞，感動詞，助動詞，助詞，（接頭語*） *活用するもの

また，意味属性情報を基準にした配列を行なうとシソーラスに準じた辞書を編集することができる。

単語体系を編集する際に，並行して意味属性別の単語一覧表を作成し，意味体系に添付する。

接辞については，接頭語，接尾語の別を明示するため，“ミス～”“～姫”のように，単語の前後に“～”を付加する。接頭語と接尾語の両方の用法がある語では“～娘～”のように表示する。

4.2 単語情報の見直し

単語意味辞書から単語体系および意味属性別の単語一覧表を予備的に編集したところ，いくつかの問題点が明らかになった。そこで，まず次の各項目について記述内容の見直しを行なった。

(1)収録単語の見直し（削除）

まず，現代国語の記述文にはほとんど使用されない古語や話し言葉(46,000 語)を単語体系の収録対象から除外した。これらは単語意味辞書構築の初期に国語辞書から収集されたものと推定される。技術文書の機械翻訳実験用に収集した専門用語(54,000 語)も単語体系の収録対象外とした。JIS 第 1，第 2 水準以外の文字コード（外字）を含む語も削除した。

また，国語辞書に準じた単語配列を行なうことから，形態素解析の必要性により単語意味辞書に登録されている語も削除した。これには，活用語のうち終止形以外の登録(2,000 語)と名詞と同形の「数詞＋助数詞」(600 語)が該当する。後者は「一部」「八戸」などであるが，人間向けには不要と判断した。

(2)収録単語の見直し（追加）

形態素解析で自動生成できることから，単語意味辞書に登録されていない語を追加した。例えば，転生名詞(1,700 語)は主に動詞の連用形を名詞化して生成されるが，人間には必要と判断した。このほか，都市名，作品名，制度名，法律名(若干)等を最新の資料を参照して追加収集した。

(3)単語体系の収録対象単語数

単語意味辞書の語数は 40 万語であるが，(1)で 10 万語強を削除，(2)1,700 語程度を追加した結果，単語体系の収録語数はほぼ 30 万語となった。

(4)意味属性の妥当性

一般名詞意味属性別単語一覧表と固有名詞意味属性別単語一覧表のそれぞれについて，同じ意味属

性内の単語を相互比較することにより適否を確認し，意味属性の修正を行なった。単語意味辞書では，単語ごとに意味属性を付与していたため，このような逆配列にしてみると，なぜこの属性なのか不明なものが多数発見された。明らかな入力ミスはともかく，ほとんどは勘違いに起因すると思われる。しかし，中には特殊な用法に対応したようなものもあり，かなり難しい作業となった。

(5)読みの見直し

単語意味辞書の前身は日本文朗読システム用として開発が開始された経緯があるため，読みは発音ベースで記述されてきた。例えば，「朗読」は「ろーどく」のように，長音は“ー”で表示されている。日英機械翻訳では読み情報はあまり重要でないため，一部未整備のままとなっていた。特に人名や地名の読みにおいて，小さい「っゃゅょ」が大きい「つやゆよ」となっているものが多数そのままになっており，これらを重点的に見直した。しかる後，長音表示を国語辞書表記へ変更した。

4.3 編集手順

単語意味辞書には様々な情報が記されているのに対し，単語体系の編集に使用される情報は単語表記，標準表記，読み，品詞，意味属性に過ぎない。また，前節で見直した情報は，必ずしも単語意味辞書の記述方針に合致していないため，入念に確認してからでなければ単語意味辞書に反映することはできない。そこで，単語意味辞書から単語体系の編集に使用する情報を取り出した“ミニ辞書”を構築し，それに見直した情報の反映を行なった後，単語体系と意味属性別の単語一覧表を作成した。

5 単語意味辞書の今後の課題

意味属性別の単語表を作ったことにより，姉妹関係にある意味属性に対照的な単語が漏れているような場合が目についた。不足語彙はすぐにわかる範囲で追加を行なったが，網羅的に収録語彙の充実を図る必要がある。

抽象語の意味属性の適否の判定は難しく，今回の見直しでは十分ではない可能性がある。意味属性別の単語表を活用して意味属性の見直しを進める必要がある。

特に意味属性の見直しの際，特殊な用法はすぐには思い浮かばないため，適否の判定に多大な時間を要したことに対する反省として，用例を付与するなどの工夫が必要であると思われる。あるいは，コーパスと密接な連携を図ることにより，単語意味辞書に新たな情報[宮崎 95]を付与する際，既存の単語に対する作業が効率化することが期待される。

最後に，辞書の宿命ではあるが，地名や時事用語などの最新語彙を常時補充していくことが必要であり，この点からも組織的に辞書を維持していく必要がある。

6 おわりに

本稿では，日英機械翻訳用の単語意味辞書の概要と，それを編集して作成した単語体系について報告した。今後は，この編集を通して明らかになった単語意味辞書の問題点の解決や，辞書とコーパスとの連携を進めていく予定である。

謝辞　日本語語彙大系の編集に関してご指導，ご討論頂いた岩波書店の宮内久男氏，上野真志氏，岡本潤氏に感謝する。

参考文献

[池原93] 池原,宮崎,横尾: 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)

[池原97] 池原,宮崎,白井,横尾,中岩,小倉,大山,林: 日本語語彙大系, 岩波書店 (1997)

[宮崎84] 宮崎正弘: 係り受け解析を用いた複合語の自動分割法, 情報処理学会論文誌, Vol.25, No.6, pp.970-979 (1984)

[宮崎93] 宮崎,池原,横尾: 複合語の構造化に基づく対訳辞書の単語結合型辞書引き, 情報処理学会論文誌, Vol.34, No.4, pp.743-754 (1993)

[宮崎95] 宮崎正弘: 辞書の記述と利用 -機械辞書の観点から-, 日本語学, Vol.14, No.4, pp.52-61, 明治書院 (1995)

[宮崎97] 宮崎,池原,横尾,白井: 日英翻訳のための意味属性体系, 電子情報通信学会, 技術研究報告, NLC97-12, pp.29-36 (1997)

[中岩97] 中岩,池原: 日英の構文的対応関係に着目した日本語用言意味属性の分類, 情報処理学会論文誌, Vol.38, No.2, pp.215-225 (1997)

[横尾97] 横尾,宮崎,池原,白井,阿部: 日英翻訳のための単語辞書, 電子情報通信学会, 技術研究報告, NLC97-13, pp.37-44 (1997)