

日本語語彙大系について

白井 諭^{*2 *1} 大山 芳史^{*1} 池原 悟^{*3} 宮崎 正弘^{*4} 横尾 昭男^{*2}

^{*1}NTTコミュニケーション科学研究所 ^{*2}ATR音声翻訳通信研究所

^{*3}鳥取大学 工学部 ^{*4}新潟大学 工学部

あらまし 日英機械翻訳における高品質な意味解析を実現するため、筆者らは日英機械翻訳システムの開発とともに、それに用いる意味辞書の構築を進めてきた。この意味辞書は、単語や表現構造の意味を体系的に分類した意味属性体系、単語に関する知識を収録した単語意味辞書、用言を核とした表現構造を収録した構文意味辞書の3つから構成される。

意味属性体系は、対象の見方や捉え方が、一般名詞意味属性、固有名詞意味属性、および、表現構造に対する用言意味属性として3種類3,000属性に分類、体系化されている。

単語意味辞書は、現代日本語の記述文への適用に耐えるよう、単語の異表記や固有名詞20万語を含む40万語に対し、文法情報のほかに、一般名詞意味属性と固有名詞意味属性が付与されている。

構文意味辞書は、現在、6,000用言に対する表現構造が日英対訳形式で16,000パターン収集され、日本語パターンの格要素の名詞に対し一般名詞意味属性を用いた制約条件が記述され、日本語パターン全体に対し用言意味属性が付与されている。

本稿では、これらの意味辞書の開発経過と、それに基づいて作成した日本語語彙大系の概要について報告する。

Introduction to "Goi-Taikei: A Japanese Lexicon"

Satoshi SHIRAI^{*2 *1}, Yoshifumi OYAMA^{*1}, Satoru IKEHARA^{*3},
Masahiro MIYAZAKI^{*4}, and Akio YOKOO^{*2}

^{*1}NTT Communication Science Laboratories,

^{*2}ATR Interpreting Telecommunications Research Laboratories,

^{*3}Faculty of Engineering, Tottori University, and

^{*4}Faculty of Engineering, Niigata University

Abstract In order to improve the quality of Japanese-to-English machine translation, we have developed a Japanese-to-English machine translation system along with the semantic dictionaries it requires. The semantic dictionaries consist of a semantic attribute system, a semantic word dictionary and a semantic valency dictionary.

The semantic attribute system gives a hierarchy of 3,000 attributes which are used to describe different characteristics of common nouns, proper nouns and predicates.

The semantic word dictionary consists of 400,000 words, including spelling variations and some 200,000 proper nouns. Words are marked with syntactic and semantic information.

The semantic valency dictionary covers 6,000 Japanese verbs, adjectives and nouns, divided into 16,000 Japanese and English pattern pairs for 6,000. The patterns are marked with verbal semantic attributes, and constraints are given on their arguments with the common noun semantic attributes.

In this paper, we describe how these dictionaries were developed, and give an outline of the published subset: "Goi-Taikei: A Japanese Lexicon".

1 はじめに

日本語彙大系[池原97]は、意味体系、単語体系、構文体系の3部から成り、日英機械翻訳用に開発されてきた意味辞書の主要な情報を人間の使用を意識して編集することにより作成された。意味体系には、一般名詞意味属性体系(2,700属性)、固有名詞意味属性体系(130属性)、文型パターン対に対する用言意味属性体系(100属性)のうちの上位36属性の各体系図と、各意味属性別の単語表が収録されている。単語体系には、一般語(10万語)と固有名詞(20万語)に対し、単語表記(異表記を含む)、読み、粗い品詞分類と、一般名詞意味属性(一般名詞の場合)または固有名詞意味属性(固有名詞の場合)が付与されている。構文体系には、用言を中心とした格構造が日英対照の文型パターン対として表現され(14,700文型)、格要素の名詞の制約条件が一般名詞意味属性により記述されている。

日本語彙大系の詳細については意味体系に収録した解説に譲り、以下では基となった日英機械翻訳用意味辞書の開発経過と日本語彙大系の編集方針について述べる。

2 日英機械翻訳用意味辞書の開発経過

本節では、意味辞書の開発経過を述べる。

2.1 日本語処理から日英機械翻訳へ

筆者らは、1980年度に自然言語処理の研究を開始した。日本文音声変換システム(JTOS)[宮崎83]の開発を最初の課題として、日本語処理の最も基本となる単語辞書と形態素解析の実現を目指した。そして、新聞記事を読み上げる実験を進める上で発生した課題、すなわち、単語辞書には一般語だけでなく人名、地名といった固有名詞が必須であること、複合語は予め網羅できず基本語分割が必要になること、固有名詞を加えると同形語の続出が問題となること、同形語を同定するには詳細な単語分類(品詞、意味、相互関係情報、等)が必要になることなどを段階的に克服してきた。

単語辞書と形態素解析の一応の完成を踏まえて、1983年度末から日英翻訳の研究を開始した。当初は翻訳支援システムの開発を課題として、構文解析、意味解析、訳語選択などの基本検討を進める予定であった。その直後、日英翻訳の商用システム1号機が発売されたことで、研究方針の見直しを余儀なくされた。改めて他機関の機械翻訳システムを再検討すると、多くのシステムでは表現と意味の対応関係をルール化することを志向し、一般語辞書を用い

て単語単位でトランスファする方式を基軸に、翻訳精度を高めるため例外処理が併用されていた。

2.2 日英機械翻訳の課題認識

当時、筆者らは言語過程説[時枝41,三浦67]に着目し、その観点から言語処理の考察を並行して進めていた。言語過程説によれば、話し手は自分の基準で対象を認識しその結果を言語として表現し、聞き手は表現を手がかりに話し手の認識の仕方を追体験しその背後にある対象を思い描こうとする。そして、この「対象-認識-表現」の関係を表現の意味とする。認識と表現との対応関係は、話し手と聞き手とで同じでなければ意味が通じないことになり、この対応関係は言語規範と呼ばれ、国語辞書に記述されている単語の語義に相当する。

また、翻訳は原言語における「対象-認識-表現」の関係を追体験し目的言語で表現し直すことであると言える。言語規範は言語ごとの文化的背景の上に成り立つものであるから、翻訳は本来的に近似であると考えべきである。これは、翻訳とは慣用表現の対応付けであるとする翻訳家の意見[中村83]により傍証される。

これを極論すると、あらゆる日本語表現と対応する英語表現を収集し、データベース検索の要領で翻訳するシステム構成が考えられる。これを工学的に実現するため、多段翻訳方式を提案し[池原87]、意味を喪失しない表現単位を網羅的に収集すること、実用文で使用される単語を網羅的に整備すること、単語の意味(語義)を体系的に分類することを課題として取り組んだ。このとき、日本語の単語や表現の語義を決定し、対応する英訳を決定する一連の手順を日英翻訳における意味解析と定義した。

意味を喪失しない表現単位としては、大小様々な表現が考えられるが、言語は開集合であるため、あらゆる表現を対象にすると網羅性を測定するのは容易ではない。石綿と荻野による「結合価から見た日本文法」と「日本語用言の結合価」([水谷83]の第2章と附録2)に着目し、体言への条件指定を詳細化するとともに日英対照形式で記述する方針で、用言を中心とする表現構造の網羅的収集を目指すことにした。表現単位を限ることにより検討対象が絞り込まれ(1万用言を想定)、使用頻度を加味することにより工学的に必要な範囲が測定可能になる(例えば、頻度で9割の網羅性、など)と考えたからである。

2.3 意味辞書の構成

意味の分類として、JTOSでは、単語を読み分けるため、一般名詞を500属性に分類し体系化してい

た。精細な訳語選択を実現するには十分でない判断し、単語の使われ方の観点から分類を詳細化する方向で見直し、一般名詞意味属性体系の構築を開始した。訳語選択では、「日本語単語＋意味属性→英語表現」の形で日英の対応付けを想定した。また、複合語の単語構成を解析する上で、固有名詞の性質を区別する必要から、一般名詞意味属性体系の一部をさらに詳細化する形で、固有名詞意味属性体系の構築を並行して進めた。一般名詞意味属性は2,700に、固有名詞意味属性は130にそれぞれ体系化された。この意味分類が日英翻訳に有効であることは実験的に示されている[池原93]。

単語意味辞書として、代表的な現代日本語の記述文である新聞文を対象とできる規模のものの整備を目標とした。JTOSで収集した一般語と固有名詞に加え、電気・通信分野の専門用語を追加収集した。表記の揺れに対応するため、標準表記を設けるとともに異表記を積極的に登録した。品詞分類はJTOSの200分類を継承し、動詞分類(意志動詞、自発動詞、等)、副詞分類(程度副詞、頻度副詞、等)などを品詞補足情報として追加記述した。また、接続属性、様相属性、時制属性を体系的に付与した。現在、単語意味辞書の基本セットとして40万語、ほかに専門語辞書として60万語を収集している。

構文意味辞書として、用言(動詞、形容詞、いわゆる形容動詞)と格要素を組み合わせた文型を結合価パターンとして日本語と英語で対にして網羅的に収集することを目標にした。格要素は名詞と格助詞からなるものを対象とし、名詞は一般名詞意味属性を用いて抽象化し、格助詞は名詞と用言の関係を示すマーカとして一般的なものを記述した。また、特定の名詞と用言が結びついて初めて意味を持つものは慣用表現として区別し収集した。初期は和英辞書から一般表現10,000パターンと慣用表現3,000パターンを収集したが、実験的に推計した結果、必要規模の半分程度しか収集されていないことが判明した[白井95]。そこで、詳細な語義分類に基づく用例[IPA87, IPA90]や作例とそれらの英訳から文型を収集する方法に移行し、現在、高頻度用言の低頻度用法や低頻度用言の文型の収集を進めている。また、収集過程を見直した結果、名詞が述語として働いているものを収集対象に加えたほか、副詞性の要素を「格要素」に、文末に様相表現がつくものを別の文型として収集するなどの収集範囲の拡張を行なった。また、文脈処理として、パターン対の相互関係を記述するため、「用言が持つ動的属性の分類」と「用言の格に対する関係」に着目した用言意味属性100属性を導入した[中岩97]。

3 日本語語彙大系の編集

本節では、日英機械翻訳用意味辞書を人間用の辞書に編集した際の検討事項を中心に述べる。

3.1 「意味体系」の編集

日英機械翻訳用意味辞書では、各意味属性は属性番号で表されており、その意味的上下関係は属性番号の値で判断される。各意味属性には名称が付与されているが、作業上の便宜に過ぎなかった。これに対して、意味体系では人間用の辞書として、意味属性の名称を十分吟味し、意味的上下関係を図的に表現した。意味的上下関係には“is-a”と“has-a”があり、そのいずれの関係であるかを視覚的に区別できるように表示した。

以上の形式で、一般名詞意味属性(2,700属性)と固有名詞意味属性(130属性)は全体を掲載した。

用言意味属性は「用言が持つ動的属性の種類」と「用言の格に対する関係」の2種類の観点から分類された100属性から成る。このうち、後者は現在も詳細検討中であり、網羅性や適切性に若干問題があるため、前者の36属性のみを掲載した。

3.2 「単語体系」の編集

単語体系への収録対象語は、現代国語文で使用される語とし、活用語のうち機械的な派生が単純でないため別登録した語や固有名詞と同形となる「数詞＋助数詞」のような語、日英翻訳の実験等の目的で導入した専門用語等は除外した。一方、単語意味辞書には派生関係にある語は派生情報を付与された基本語のみを収録し、解析過程で一定の条件を満たせば派生語を生成している。そのような基本語のうち動詞転生型の名詞を派生させて収録した。合計30万語である。

各単語は、単語表記、異表記、読み、品詞、意味属性を表示した。品詞は、意味辞書で用いられている200種類を21種類に縮退させた。意味属性は、固有名詞意味属性は固有名詞にのみ付与されているが一般名詞意味属性と一定の対応関係があることから、一般名詞は一般名詞意味属性を、固有名詞は固有名詞意味属性を掲載した。

3.3 「構文体系」の編集

構文体系には、構文意味辞書に収録されている6,000用言16,500文型のうち、翻訳実験等で検証済みのものを中心に再吟味し、最終的に14,700文型を収録した。各文型は、日本語文型、英語文型、文型パターン対の用言意味属性、格要素への制約条件、英語文型の動詞の進行形・受動態可否の5

種類の情報を掲載することとし、一覧性を考慮して日本語文型と英語文型を一次元化して表示した。

文型には一般表現文型と慣用表現文型があり、一般表現文型では原則としてすべての格要素が一般名詞意味属性による制約条件で指定されるのに対し、慣用表現文型では1つ以上格の名詞が字面で指定される等の違いがある。形式的には両者は同一の形式であるので、構文体系では表現上は区別せずに掲載した。

構文意味辞書には、慣用表現の名詞に対する修飾の可否(例えば、「油を売る」の「油」は修飾不可)や英語文型の要素の文法情報(主語、目的語、冠詞、等)が記述されている。これらを文型の一次元表示に加えると見づらくなるため、構文体系では割愛した。構文意味辞書には、日本語文型の用言、英語文型の動詞に対する否定指定などの条件は属性コードで指定されている。人間の使用を考え、これらの属性コードを代表的な記述(否定→「ない」、状態→「ている」、等)で代用した。

3.4 その他の情報の編集

意味属性別単語表として、意味体系に収録した3種類の意味属性体系を構成する意味属性から見た日本語単語と文型パターン対を一覧表にまとめ、意味体系に併録した。また、構文体系に収録した文型パターン対の検索の便宜を考えて、日本語名詞索引として該当する名詞を持つ慣用表現文型を、英語索引として英語文型に含まれる単語を、それぞれ一覧表としてまとめ、構文体系に併録した。

4 おわりに

日英機械翻訳用の意味辞書の開発の背景と、意味辞書に基づいて人間用に編集した日本語語彙大系の編集の概要を報告した。本プロジェクトでは様々な試行錯誤を通し、言語処理を進める上での課題を明確化するとともに、それに必要な言語知識の編集の第一歩を記すことができたと考えている。その一方で、特に今回の日本語語彙大系の編集を通して、未解決の課題が多いことも分かってきた。

今後の課題としては次のようなものに取り組む予定である。意味体系として、抽象語に対する細分類が必要であり、木構造の一部は修正する方がよさそうである。単語体系として、新語への対応が不可欠であり、単語の相互関係の記述を強化する。構文体系として、おそらく25,000文型くらいを収集する必要があり、またそれらの検証を容易にするため文型ごとに用例を付与しておく。

謝辞 本辞書プロジェクトの推進には、日本語語彙大系の共編者である小倉健太郎、中岩浩巳、林良彦の各氏ほか、奥雅博、内野一、松尾義博、Francis Bondの各氏らのご協力を得た。実務の面では、細井純子、松吉久美子、成田恵理、阿部さつき、八木晶子、井上浩子、渡邊いづみ、関嘉代の各氏を始めとする方々に長期にわたり多大なご協力を頂いた。日本語語彙大系の編集、出版に際しては、岩波書店の宮内久男氏、上野真志氏、岡本潤氏らのご指導を頂いた。また、長尾眞京都大学教授(現総長)、田中穂積東京工業大学教授には推薦文を寄せて頂いた。以上の方々に深謝する。

本稿執筆中の1998年10月18日に宮内久男氏が急逝されました。謹んでご冥福をお祈りします。

参考文献

- [池原87] 池原,宮崎,白井,横尾: 言語における話者の認識と多段翻訳方式, 情報処理学会論文誌, Vol.28, No.12, pp.1269-1279 (1987)
- [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)
- [IPA87] 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 & 辞書編 (1987)
- [IPA90] 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 & 辞書編 (1990)
- [水谷83] 水谷,石綿,荻野,賀来,草薙,青山: 文法と意味I (朝倉日本語新講座3), 朝倉書店 (1983)
- [三浦67] 三浦つとむ: 認識と言語の理論I~III, 勁草書房 (1967)
- [宮崎83] M. Miyazaki, S. Goto, Y. Ooyama, & S. Shirai: Linguistic processing in a Japanese text to speech system, ICTP '83, Tokyo, pp.315-320 (1983)
- [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)
- [中村83] 中村保男: 翻訳はどこまで可能か, ジャパンタイムズ (1983)
- [白井95] S. Shirai, S. Ikehara, A. Yokoo, & H. Inoue: The quantity of valency pattern pairs required for Japanese to English MT and their compilation, NLPRS '95, Seoul, Vol.1, pp.443-448 (1995)
- [時枝41] 時枝誠記: 国語学原論, 岩波書店 (1941)

付録(日本語語彙大系の組方見本)

以下、『意味体系』所収の一般名詞意味属性体系、『意味体系』所収の意味属性別単語表、『単語体系』、『構文体系』の見本を順に示す。

