ＩＰＡにおけるテキストコーパスの作成

3Ｈ－1

橋本三奈子

情報処理振興事業協会(IPA) 技術センター

1　はじめに

ＩＰＡ技術センターでは、動詞・形容詞・名詞辞書を作成する過程で見出し語の分析のために新聞、教科書、文芸作品などの実例をデータベース化してきた。けれども単語分割や品詞情報付加などの手を加えていないため、検索効率が悪い。そこで今回、一部のテキストに対し、単語分割と品詞情報付加を行なうことにした。

欧米に比べ、日本ではコーパスと呼ばれるような電子化された言語データの作成および共有化が遅れている。ＩＰＡのような公共機関が、電子化データを公開することが望ましい。そこで、単語に分割し、品詞情報を付加したデータベースを、「ＩＰＡコーパス」と名付け、一般公開することにした。

当稿では、ＩＰＡコーパスに収録するテキストや、付加する品詞情報について報告する。

2　辞書記述におけるコーパスの必要性

ＩＰＡ技術センターでは、動詞・形容詞・名詞辞書を作成する過程で見出し語の分析のために新聞、教科書、文芸作品などの実例約４３万文をデータベース化してきた。これまでの研究成果である「計算機用日本語基本動詞辞書」「計算機用日本語基本形容詞辞書」「計算機用日本語基本名詞辞書」は、新聞・教科書・論文・マニュアルなどのコーパスから得たデータを分析して抽出した情報を記載した結果である。

ただし、これまでのコーパスは、テキスト名や文番号を付与しただけのデータであり、言語情報（タグ）はつけられていない。したがって、ひらがなで表記された見出しなどを検索するとかなり無駄なものが入ってしまう。例えば、実例４３万文中で「かに」は5353個あり、そのうち名詞の「かに」（蟹）はわずかに５個だった。他は、「確かに」「明らかに」「どう進めるかについて」など「かに」を含む他品詞語の一部であった。実例が大量になればなるほど、このような検索結果から必要な単語だけをとりだして分析するのは効率が悪い。そのため、単語分割され、品詞が付与されたコーパスを作成することが急務となった。

Building a Corpus at IPA
Minako HASHIMOTO.
Software Technology Center,
Information-technology Promotion Agency, Japan
3-1-38 Shiba-koen, Minato-ku, Tokyo, 105 JAPAN

3　収録するテキスト

これまでに収集したテキストは、公開するにあたって著作権などの問題があるため、改めてテキストを収集することにした。今回、大きく分類して次の五つのタイプの文章を収録した。

一つは、公開済みのＩＰＡＬ動詞辞書、形容詞辞書、名詞辞書に収められている意味記述文および文例であり、約１５０００文である。これらは、ＩＰＡＬ辞書を利用する際に有用であると同時に、基本語の意味・用法をある程度網羅したものであるため、品詞情報が付加されたものがあれば、形態素解析や構文解析の評価のための例文集としても適切である。

二つ目は、『日本語表現文型 中級』（筑波大学日本語教育研究会,凡人社,1983)の中に収められている、「文型・文法」欄の例文約１６００文である。これは、日本の大学あるいは大学院に留学する一般外国人留学生を対象とした中級程度の日本語教材であり、一般的に日本の大学で要求される理解・表現の型が分類・整理されているものである。どの例文も基本的な文型に沿ったものであるため、これも形態素解析や構文解析の評価のための例文集としても適切であると考えられる。

三つ目は、岩波ジュニア新書７冊の文章であり、約１３０００文ある。中学生・高校生を対象に科学的な問題を平易に解説した文章である。構文構造、意味構造さらには文脈構造の分析にも有益なテキストとなるであろう。

四つ目は、岩波新書１３冊分の文章で、約２８０００文ある。文化・科学・政治等の問題をその道の専門家が広く一般の読者のために書き起こした論述文である。先にあげた岩波ジュニア新書よりも文章の難易度は高い。岩波新書は、その内容や知名度からも日本語のテキストコーパスとして適していると思われる。

最後に、大学入試問題の現代国語の出題文６４テキスト中の約２７００文である。大学入試問題を取り上げたのはそれぞれが日本の著名な作家の著述物であり、長い作品の一部分とはいえ、論旨の展開が見られ、内容豊富なものである、と判断したからである。文芸作品もあり、修辞的な表現も含まれている。

以上、約６００００文を収録した。比較的入手しやすい新聞記事などは避け、一般には入手が難しく、文章の質の高いものを収集するよう、努力した。難易度的にもバランスのとれたものになっていると考えられる。

4　タグセット（品詞体系）

品詞体系の作成や、それぞれの単語についての品詞の認定は非常に困難な問題である。単語に付与した品詞には、主観を含まざるを得ない。利用者が自分の研究目的に合わせて自由に取捨選択あるいは変更して利用できるように、ということを心がけてタグセットを作成した。基にした文法や用語なども統一しきれていないところも多いが、特に、学校文法の扱いと異なるような箇所には括弧つきで学校文法の品詞を付加しておくなど、利用者のカスタマイズが簡便に行えるように考慮してある。

品詞については、第１レベルとして、次のものをたてた。

(1)　名詞
(2)　動詞
(3)　形容詞
(4)　形容動詞
(5)　副詞
(6)　連体詞
(7)　接続詞
(8)　助詞
(9)　助動詞
(10)　感動詞
(11)　記号
(12)　その他

これらの第１レベル１２種の品詞に対して、必要に応じ第５レベルまで、言語学的に妥当で、かつ言語処理に有用と思われるような分類を加えてある。

日本語はわかち書きされていないので、語の認定が非常に難しい。一番扱いに困ったものは、接頭語、接尾語といわれる類である。例えば、「甘さ」は「甘」と「さ」に分割するよりも「甘さ」として一語の名詞として認める方が望ましい。しかし、「恐いもの見たさ」の「さ」を、直前に現われる「たい」の語幹「た」につなげると「たさ」となるが、これを一語とするのは誤りである。また、「県に認可申請を提出済みだ」では、「済み」を直前の「提出」とつなげて名詞としてしまうと、その前に現われる二格やヲ格の扱いが明確でなくなる。構文的には、「恐いもの見たさ」「県に認可申請を提出済み」を一つの語として扱うべきだが、これは、構文解析まで行なわなければ解析できない。そのため今回は、その語が他の語と結び付いて品詞を与えることのできる一つの語になることを明示するため、「名詞」「形容詞」等の下位分類として「接頭」「接尾」を認めた。

品詞の付与にあたっては、一般公開されている形態素解析システムJUMAN のエンジン、および同じく公開されている「汎用日本語形態素解析規則」の辞書と規則を用いた。その出力に対して、複数の形態素をつなげ、タグを付与するツールを作成した。この結果を人手で見直し、最終的に「正解」と判断されたものが記載される。

ところで、正しい品詞の決定には意味や文脈を必要とする。例えば、「テレビを見るがいい」における「が」は接続助詞（「テレビを見るが、いいですか」の意）と格助詞（「（そんなに言うのなら）テレビを（好きなだけ）見るがいい」の意）とが考えられる。この判断には、文脈の解析が必要となる。今回作成したタグセットは、人手による見直しの工程を重視したものである。完全に形態素解析レベルで付与できるタグセットを考えるならば、また別のものを用意するべきである。

5　おわりに

計算機用辞書を作成するための言語資料として岩波新書をはじめとする実例文章を電子化し、単語に分割して品詞情報を付加した。品詞情報が付加されているため、「ある」「なる」など、ひらがなで表記されることが多く、他品詞語とまぎれやすい動詞などに対する検索効率が上がった。今後は係り受け情報を付加し、動詞などの文型の記述に役立てたい。

［謝辞］データの利用を許可してくださった筑波大学日本語教育研究会、岩波書店および長尾真先生に感謝の意を表する。

［参考文献］橋本三奈子，荻野紫穂，徳永健伸，元吉文男，井佐原均「ＩＰＡコーパスの概要」『ＩＰＡＬシンポジウム '95』情報処理振興事業協会(1995)