

日本語品詞列集成の作成

西村 恕彦 萩野 綱男

(電総研) (東京大学)

1. 品詞列集成とは

国立国語研究所は、新聞の語彙調査に用いたデータ(1/3)を公開した。これは原文テープと辞書テープとからなる。これを利用して、新聞の文章中に現われた各単語に品詞を与え、その品詞の並び(品詞列)を整理し、それに度数と文例を添えた資料集を作成した。これが「日本語品詞列集成」である。簡単にいえば、品詞K W I O索引である。右順篇903ページはすでに完成し、左順篇902ページと名詞篇247ページが現在準備中である。

右順篇のページの例を表1に示す。左から順に、一連番号、12品詞列、出典番号、文例である。一連番号は、約98万個の文例すべてにつけた通し番号である。12品詞列は、3つの部分からなる。左文脈5品詞、見出し5品詞、右文脈2品詞である。見出しの長さ5の品詞列パターンが変わったところで、1例ずつ文例を添え、1行に表示してある。だから一連番号がとびとびについているのである。

一連番号は、引き算によつてある品詞列パターンの出現度数がわかるようになっている。たとえば表1のいちばん上の2行にある一連番号を引き算する。

$$594825 - 594822 = 3$$

これは、

数詞 * 付属語 * 格助詞 * 動詞 * 動詞

の出現度数が3であることを示す。

また、長さ4の品詞列、

名詞 * 係助詞 * 動詞 * 動詞

の出現度数は、表1の上半分あたりの一連番号を引き算し、

$$594883 - 594862 = 21$$

となる。

出典番号は、国研が新聞紙面上の標本抽出区につけた識別情報であつて、たとえばA0は朝日新聞朝刊を表わす。

文例は、短単位ごとに分かち書きされている。ページの右の方に、縦に空白の線が通つてみえる。ここが見出しと右文脈の境界である。ここから左右に順に見てゆけば、品詞列中の各品詞と、文例中の各語との対応がつけられる(表2)。

行の順序は、ちよつと特別な配置になつてゐる。長さ12の品詞列の各語を左から順に、第1語、第2語、…、第12語とすると、第10語を第1ソートキー、第9語を第2ソートキー、…とした(表3)。これを右順のソートとよび、それにしたがつてこの品詞列集成を右順篇と名付けた。

品詞は22に分類されていて、ソートにおけるその配列順は、後出の表4に示すとおりである。

表 1

[illegible]

表 2 文例と品詞列の対応

い	た	女	店員	五	人	が	焼け	死ん	だ
動	陳	付	名	数	付	格	動	動	陳
動詞	助動詞	付属語	名詞	数詞	付属語	格助詞	動詞	動詞	助動詞

表 3 ソートキーの位置

語番号	1	2	3	4	5	6	7	8	9	10	11	12
品詞列	終	動	陳	付	名	数	付	格	動	動	陳	不
ソートキー	10	9	8	7	6	5	4	3	2	1	11	12

2. 作成上の問題点

今後、類似の作業を行なう人のために、二、三の問題点をあげておく。

2.1 手順の概要

- (1) 原文テープの各長単位に、1～76万の通し番号をつける。
- (2) 各長単位を切り出し、漢字符号の順にソートする。
- (3) 同じ順に並んでいる辞書テープと照合し、短単位に切り、品詞を与える。
- (4) もとの通し番号順にソートし、原文の順序を復元する。
- (5) 長さ12の品詞列を切り出す。
- (6) 文例のついた品詞列を、表4の品詞コードの順にソートする。

2.2 大量データ処理

同じデータを用いた植村の先例の報告にもあるとおり、長単位76万＝短単位98万＝154万字のデータを処理する作業は大変である。一つのファイルが732mの磁気テープ8本にもなると、取扱い操作の便を充分に考えなければならない。処理した磁気テープは、全部で57本に及ぶ。

ソートに要する時間も長い。文例のソートは、大形機を占有して3時間かかった。この機械は本来TSS用なのであるが、TSS環境のもとで大量データのソートを行なうことは、不可能といつてよい。

品詞列集成の出力印字量も膨大である。ここでは1割弱を抜き出して900ページの本の形にしたが、もしすべてを印刷すれば、百科事典級の印刷物になる。

プログラムは、70本、約15000行に達した。プログラム作成にあたっては、オンライン端末によるエジタの全面的な活用が必須だった。プログラム言語は、ドキュメンテーションと効率の面から、主にCOBOLを利用した。

これらの作業の全体を通じて、大形計算機設備が有効であつたことは、以上のとおり明らかであろう。

2.3 符号変換

HITACからTOSBACへの磁気テープの変換と、漢字符号の変換との両方が必要だった。特に後者は、国研漢テレ符号と電総研の漢字プリンタ(昭和情報機器製のT4100)符号の2系だけを用いたが、いずれも符号表やドキュメントが不十分で、この点にも思わぬ手数がかった。

2.4 多品詞語の処理

国研の辞書テープは、長単位を見出しとして、その短単位への区切り方と、各短単位の品詞が記載されている。ところがこの品詞は、新聞の語彙調査で同語異語を判定するなどの目的でつけられたものであるため、あまり精度が高くない。

さらに品詞列集成の作成にあたって困るのは、1つの短単位に最高5個までの品詞

表 4 品 詞 一 覧

記号	語 数	名 称	国 研 コード	多 品 詞 語	注
1 空	5 2 2 2 8	空品詞			標本抽出区の外側
2 名	3 0 2 0 6 3	名詞	1, 2, 5, 8, 9	名+陳(ん)	純名詞, 連用形転成, 形容詞転成, 固有名詞, 代名詞
3 サ	7 7 0 1	サ変動詞語幹	3		続行, 期待, ひとり立ち, マッチ
4 ナ	3 4 4 9	形容動詞語幹	4	同じ	明らか, 有力, はるか, スマート
5 付	7 1 4 8 3	付属語(接辞)	6		通り, 的, 日, 新, 全, 第, ご, キロ
6 数	4 2 4 1 5	数詞	7		一, 二日, 半, 四十, いくら, セブン
7 ア	1 4 1 6 5 9	数字(算用, ローマ)	X		1, 2, 3
8 連	2 0 4 8	連体詞	D		大きな, どの, こんな
9 副	7 2 8 6	副詞	C	いま, こり, いずれ	なお, まだ, そろそろ, 最も
10 接	3 7 4 0	接統詞・感動詞	A, B	また, ただ, もっとも	しかし, および, はい, 有難う, では
11 動	5 1 1 6 7	動詞	E	連+動(ある)	求める, する, 行わ, 始め
12 用	5 2 4 4	動詞類	+	名+動(考え)	始める, 続け, おり, 動詞性接辞
13 形	5 6 9 1	形容詞	L, -	形+陳(ない)	少ない, 新しく, よい, やすい, 形容詞性接辞
14 陳	2 9 7 0 4	助動詞	P		ます, た, で, う
15 格	9 9 3 0 2	格助詞	R		が, の, に, を, と, で
16 係	2 2 5 8 4	係助詞・副助詞	R		は, も, まです, しか
17 終	1 5 2 6 3	終助詞・接統助詞	R		か, ね, て, ながら, ば, ので
18 記	8 5 7 9 8	記号・符号	Y		f, !, {, .m², ①, ㊦
19 読	3 0 0 8 3	読点(コンマ)	Y		読点, コンマ
20 句	1 5 2 7 7	句点(小数点)	Y		句点, ビリオド, 小数点
21 文	3 5 1 2 3	文区切り			
22 不	7 3 5 7	無関係・不明	Z	名+助詞	振仮名, ドゴール仏大統領も, the

がつけられていることである。この多品詞短単位の語数は、異なりは2209と少ないが、延べは156271と16%に達する。品詞をキーとしてソートし配列するのに、多品詞のままでは、どうしてもしょうがない。そこでこの16%をいろいろな手法で、およそ1%にまで下げた。

2.4 / 辞書の修正

/長単位が2つ以上の短単位に切れる場合には、品詞が一意にきまることが多い。これは異なりで約1/8万語ある。2品詞の組合せになるものは34語であり、3品詞以上になるものはなかった。逆に、多品詞語のほとんどは、/長単位=/短単位のものである。

これを個別に検討してみると、辞書の間違いであるかまたは用例がきわめて稀であると想像されるものが多かった。これを植村の用例索引マイクロフィッシュで調べ、辞書を修正した。これによつて、多品詞短単位の延べ語数のおよそ9割が1品詞になった。

辞書修正の例

ます	{	助動詞	1041	}	助動詞にする
		名詞(魚)	1		
		動詞(増す)	0		

2.4.2 判別文法

辞書修正後も、異なり642、延べ16000の多品詞短単位が残った。/長単位が2短単位以上からなる場合に、多品詞にならないのは、文脈の効果によると思われるので、同様に多品詞短単位の左右の文脈を評価すれば、1品詞になるかもしれない。こういう前提のもとに、多品詞の組合せ47種に対して、223個の判別文法を作った。

判別文法の例 (? は名詞と動詞の多品詞)

? * 助動詞だ	→	名詞	わけ * だ
? * 助動詞	→	動詞	わけ * た
? * 接続助詞	→	動詞	わけ * ながら
? * 助詞	→	名詞	わけ * は
助詞の * ?	→	名詞	の * わけ
助詞 * ?	→	動詞	に * わけ

判別文法は予期したほど有効ではなかったが、それでも延べ5000ほどの短単位が1品詞になった。そして残ったものは、異なり398、延べ9856の多品詞短単位である。これらは無理にどれか1つの品詞に押し込んだ。

2.5 品詞分類

国研の品詞分類を基礎にして、結局は表4のような系に落ち着いたが、そこに到るまでには多くの試行を重ねた。国研の分類からの変更の規準は、構文的な調査の目的にかなう有意味な情報が取り出せるようにすることであつて、(1)純名詞・固有名詞・代名詞や、語数のあまりに少ない品詞を統合し、(2)不明・情報無視、記号・符号、助詞の類を細分した。

すなわち、接続詞と感動詞は統合され、文区切り、句点、読点が固有の品詞として独立した。助詞は3分された。

3. 品詞列集成の利用・分析の例

品詞列パターンの出現度数は、一連番号の引き算によつて求めることができるが、長さ3以下の品詞列については、別に集計表が作つてあるので、直接に延べ度数を知ることができる。以下に、これを用いて日本語構文の分析を行なう例を示す。

(1) 出現度数の高い句

長さ3の品詞列のうち出現度数の高いものを示すと、表5のようになる。これを見ると、数字と記号の混じつた列と、((名詞)ⁿ * 格助詞)ⁿ * 動詞という2つの典型的なパターンが現われる。前者は数値の表現のパターンであつて、広告欄などに多い。後者は日本語の構文の基本となる形である。

(2) マルコフ連鎖

文区切りから出発して、その前(左)に何が来やすいかを順次にみてゆく。同じパターンを2度とらないようにすると、次のような連鎖が構成される。

... * 名詞 * 付属語 * 格助詞 * 名詞 * 名詞 * 格助詞 * 動詞 * 助動詞 * 句点 * 文区切り
これが日本語の文のモデルとして、もつともらしく見えることは、明らかであろう。

(3) 名詞句

名詞を中心語とする名詞句について、それに先行する修飾語を調べる。

名詞 * 名詞	104472
名詞 * 名詞 * 名詞	43672
連体詞 * 名詞	1662
読点 * 連体詞 * 名詞	552
動詞 * 名詞	8705
格助詞 * 動詞 * 名詞	4680
形容詞 * 名詞	2168
格助詞 * 形容詞 * 名詞	704

名詞に先行する修飾語は、名詞、連体詞、動詞、形容詞などである、ところがそれぞれの修飾語のさらに左にくる語は、修飾する品詞ごとにまったく様子が違う。名詞が修飾語である場合には、さらに名詞が先行することがいちばん多く、これらは複合語になると考えられる。ところが連体詞が修飾語である場合には、先行するものは読点であり、ここに句構造の大きい切れ目のあることが想像される。

動詞や形容詞が修飾語である場合には、格助詞が先行する。もつと長い文脈について修飾関係をみなければならぬ。

表5 長さ3の品詞列

数字 * 数字 * 数字	45746
名詞 * 名詞 * 名詞	43672
名詞 * 格助詞 * 名詞	27869
名詞 * 格助詞 * 動詞	18670
数字 * 数字 * 記号	17235
名詞 * 記号 * 名詞	17210
記号 * 数字 * 数字	16482
格助詞 * 名詞 * 格助詞	16126
名詞 * 名詞 * 格助詞	15647
数字 * 記号 * 数字	15194

4 品詞列集成の問題点

品詞列集成の出力結果を調べてゆくと、二つの問題点に気付く。

一つは、品詞のつけ間違いおよび品詞不明の類がわりに多いことである。両者を合わせて / % を下ることはないと思われ、特にこの品詞列集成は、/ パターンごとに文例を選出表示しているので、この数字は小さいものではない。

作業にあたつて、何度も品詞の修正を行なつたが、なお完全というには遠い。多品詞や不明の類もさることながら、もともと国研のデータがこのような目的で作られたものではなかつたのだから、やむをえない面もある。しかしまた、そのような検討のためにも、この資料が役立つだろう。

もう一つは、品詞列パターンを分類・配列する単位としては、短単位が短かすぎるということである。文の構文的な構造を調べるのに、短単位の担っている機能が小さすぎるのである。名詞の連続を / つの名詞句とみなすとか、サ変語幹 * 動詞スルを / つの動詞とみなすなどの処置をすれば、もう少し構文的な目的にかなうだろう。

現状では、直接的には語構成の研究に寄与する面が大きいと思われる。現在準備中の名詞篇は、3 個以上の名詞の連続をすべて表示したものであつて、名詞や付属語に関するこの面の研究に役立つ。

最後に、多品詞語や品詞分類その他の多くの方面で、渡辺孝野氏（計量計画研）のお世話になつた。また、植村俊亮（電総研）、真子ユリ子（電総研）、谷津直和（当時電通大）の諸氏の援助にも感謝したい。