3B−7

計算機による日本語文書校正のための基礎データの収集と解析

その2　日本語解析のための辞書の作成

鈴木　恵美子　　中山　妙子　　吉岡　千草　　山田　圭以　　福田　恭子
東京家政学院筑波短期大学

1. はじめに

我々は、将来的には計算機により、日本語文書の校正処理を行うことを目的として研究を進めている。特にワードプロセッサで作成された文章にどのような誤りがあるのかに注目して調査した。

また、ある特定の表現の前後の単語に関係があるかを調べるために、前後の単語に詳しい情報を与えるための辞書データベースを構築しようとしているのでそれについても述べる。

2. ワードプロセッサを使用した文章に現れた誤り

ワードプロセッサによって作成された文章には、どのような誤りがあるかを知るために学生の作成したレポートを調査した。その調査結果（表1）と書き間違いの例（表2）を挙げる。

表1　ワードプロセッサによって作成された文書における誤りの調査結果

分類	項目	合計	全体の%
スタイル	文の長さ	76	8.3
	自立語の長さ	15	1.6
	付属語の長さ	2	0.2
	接続詞の使用	14	1.5
	受け身の使用	31	3.4
	表記のゆれ	111	12.2
	助詞の使い方	82	8.9
	漢数字とアラビア数字	0	0
	かなとアルファベット	2	0.2
	言葉使い	196	21.3
	慣用句の使い方	4	0.4
	その他	87	9.5
ミスタイプ	句点の使い方	54	5.9
	読点の使い方	28	3.0
	ローマ字かな入力	85	9.2
	ひらがな列	10	1.1
	かっこの対応	8	0.9
誤変換	同音異義語	24	2.6
	未変換の漢字接尾語	82	8.9
	助詞の誤変換	4	0.4
	漢字複合語中の誤り	4	0.4
合計		919	99.9

調べているときに明らかにかな漢字変換のミスとわかるものが目についた。この誤りは手書きのときにはない誤りなので、ワードプロセッサを使用して文章を作成するときには気をつけなければならない。

表2　書き間違いの例

誤	正
またわ、	または、
ネットワークステム	ネットワークシステム
割り当てらる	割り当てられる
サーバー	サーバ
RISC	RISK
デファクト	ディファクト
データーショウ	データショウ
始めて	初めて
LAM	LAN
セキュリィ	セキュリティ

3. 辞書データベースの構築

我々は辞書データベースを構築するにあたって、単語見出し、品詞、活用形、活用語尾、接続情報という5つの項目を作成することにした。方法として、市販されている三省堂の新明解国語辞典を参考に必要な項目だけを抽出したデータ（表3）を作成することにした。

表3　新明解国語辞典から抽出したデータ

```
０００００１０　@あ@（亜・阿・唖・鴉）　「｛漢語の造語成分｝§
０００００２０　@ア@§
０００００３０　一アジアの略。「中央―３」§
０００００４０　二アフリカの略。「南ナン―１」§
０００００５０　三（日本）アルプスの略。「南ナン―連峰」〔一は亜、二は阿とも書く〕§
０００００６０　@あ@１（感）§
０００００７０　一呼びかけの声。「一君、ちよつと」§
０００００８０　二急に・驚き（思い出し）などした時に出す、言葉にならない言葉。〔一は「ああ」、二は「ああ」「あつ」とも§
０００００９０　言う〕§
```

Colection of data and its analysis to be used for japanese proofreading by using the computer
emiko SUZUKI, taeko NAKAYAMA, chigusa YOSHIOKA, kei YAMADA, yasuko FUKUDA
Tokyo Kaseigakuin Tsukuba Junior College

本研究は、財団法人日本科学協会の笹川科学研究助成によって実施したものです。

しかし、新明解国語辞典はデータが最後までそろわなかったために日本電子化研究所のＥＤＲ辞書を参考に作成することにした。

Net Ware上にネットワークドライブを開設し、ＥＤＲ辞書の磁気テープをダウンロードし、データとした。

ＥＤＲ辞書には日本語単語見出し、概念ＩＤ，英語概念見出し、日本語概念見出しの４つの項目があったので、単語見出しと概念ＩＤだけを抽出した（表４参照）。「概念ＩＤ」とは「単語見出し」を区別する英数字である。

表４　ＥＤＲ辞書の見出しと概念ＩＤ

```
06お 祭 り :e8027
06お 祭 り :e8028
06お 祭 り :e8029
06お 祭 り :3BD8CB
12お 祭 り さ わ ぎ :3CE736
12お 祭 り さ わ ぎ :3CFC36
10お 祭 り 騒 ぎ :3CE736
10お 祭 り 騒 ぎ :3CFC36
08お 祭 騒 ぎ :3CE736
08お 祭 騒 ぎ :3CFC36
04お 斎 :0E7C13
04お 菜 :3BED95
08お 菜 好 み :e739e
08お 菜 好 み :e739f
06お 菜 箸 :e76c5
06お 作 り :e7b79
06お 作 り :3C1B36
04お 匙 :0E76F3
04お 匙 :e76f4
04お 匙 :e76f5
04お 匙 :bb62d
04お 札 :3CF925
04お 薩 :e76fa
04お 三 :e7751
04お 三 :e7752
04お 三 :bb62f
```

新明解国語辞典とＥＤＲ辞書の項目の抽出方法は同じである。ここで、私達が行った辞書項目抽出手順を述べたいと思う。

１．どのような項目がデータとして入力されているのかを調べる。

２．どの項目を抽出するかを決める。

３．項目を抽出するプログラムを作成する。
①プログラムを作成するために、入力ファイルの形態を調べる。
②入力ファイルの容量が大きすぎるために画面に出力できなかったので、任意の範囲を別ファイルにコピーするプログラムを作成する。
③②で作成したプログラムを使用して入力形態を調べる。
④抽出したい項目の前後の文字をそれぞれキーワードにして、入力ファイルを１レコードずつ読み込み、項目の始まりのキーワードを探す。
⑤キーワードが見つかった時点で、辞書項目を抽出し始める。
⑥⑤で終わりのキーワードを見つけたら、キーワードの直前までの文字列を辞書項目ファイルに書き込む。
⑦④～⑥までを入力ファイルの最後まで繰り返す。

４．必要な項目を抽出したデータができる。

4. ＥＤＲ辞書の単語見出しについて

データベースを構築するにあたって各項目ごとのバイト数を決める必要があった。そのため、参考としてＥＤＲ辞書の単語見出しの最大文字数を調べ、文字数別の単語見出しの数を数えた。

その結果、最大文字数は36文字あることがわかったので、我々が構築する辞書データベースの「単語見出し」のバイト数も決まった。

単語見出しの総数は431,028個であった。その中で最も多いのが２文字の単語見出しで全体の約４分の１にあたる107,630個あった。全体的に見ると文字数が多くなるにつれ、見出しの数が少なくなっていることがわかる。

表５　単語見出しの文字数別の見出しの数

文字数	見出しの数	文字数	見出しの数
１文字	11,616 個	１９文字	1,075 個
2	107,630	２０	906
3	87,683	２１	738
4	87,366	２２	540
5	40,850	２３	355
6	24,465	２４	234
7	14,233	２５	153
8	11,479	２６	101
9	8,883	２７	70
１０	7,500	２８	57
１１	6,085	２９	52
１２	4,824	３０	22
１３	3,955	３１	6
１４	3,137	３２	1
１５	2,426	３３	1
１６	1,814	３４	0
１７	1,496	３５	0
１８	1,273	３６	2

5. おわりに

今後の課題としては、「概念ＩＤ」の英数字はどのように「単語見出し」を区別しているかを知るためにソートし、「概念ＩＤ」を参考にして、「単語見出し」の意味素性を「接続情報」の項目にしたい。

その他の項目（品詞、活用形、活用語尾）は別に調べてデータベース化する予定である。

【参考文献】

［１］高井ほか："概念体系の開発思想"，第41回情報処理学会全国大会予稿集，７Ｓ－４

［２］岸本ほか："概念分類項目の設定"，第41回情報処理学会全国大会予稿集，７Ｓ－５

［３］横田ほか："概念分類項目の関係記述"，第41回情報処理学会全国大会予稿集，７Ｓ－６

［４］Borland International, Inc.
「Turbo Pascal　リファレンスガイド」
㈱　マイクロソフトウェア　アンシェイツ