校正支援における形態素の認定

5E-2

清水 富門　　役田 敬子　　披田野 陽一　　柏木 貞一
(株)日本電子化辞書研究所

1．はじめに

我々は、当社で開発中の単語辞書、概念辞書の実証評価を目的とする日本語文校正支援システムを開発中である。同校正支援システムの第一のステップは、文中で使われた形態素の認定である。形態素の認定を基に文の構文解析、意味解析がなされ、校正支援の各項目が正される。従って、形態素の認定精度は、校正支援システムの性能を決定する重要な尺度である。校正支援システムは、その目的から考えて、理想的に表記された日本語文が期待できない場合に用いられるもので、ひらがな、漢字の使い分けの不適切な表記、漢字や送り仮名の誤った表記に関しても、その素性を認定できなければならない。本報告では、様々な問題を含んだ、一般の日本語文について、文中の形態素を精度よく認定し、校正支援を効果的に行う方法について検討した結果を述べる。

2．　日本語文中の形態素の認定

ここでは、我々が考案した形態素認定方法の主な特徴をいくつかの項目に分けて述べる。

(1)　形態素のテーブルを単語辞書とは別に持ったこと．
我々のシステムは、書き言葉の"ゆれ"る部分を処理の対象としているので、その表記上の単位を形態素と定義している。従って我々が形態素と呼ぶものは、意味的、機能的なまとまりを持った通常の語とは異なる。例えば、五段活用の動詞の語幹('刺'等)やその語尾の変化する形('さ','し','す','せ','そ'等)は形態素として形態素テーブルに登録される。また、"本棚"という語は、原理的には、'本棚','ほん棚','本だな','ほんだな'と４通りの表記が可能であるが、いづれの表記も形態素テーブルに登録することができる。(実際に登録するかどうかは我々が校正の対象とする日本語文の中で上記の表記が実際に使われることがあるかどうかを考慮して決めてゆく。)

また、"収獲"は、"収穫"の誤表記として生じ易い例であるが、こういったものも形態素テーブルに登録することができる。また、"あらわす"という言葉は漢字で書くと、"表す","表わす","著す","著わす","現す","現わす"とさまざまに表記されるが、形態素テーブルには、"あらわ","表","表わ","著","著わ","現","現わ"のすべてが登録されることになる。我々のシステムは、この様な形態素テーブルを用いて、文中に表われた表記をまず認識し、それから文中で使われた言葉の認定を行う。表記から、それをもたらした言葉を導くためのインデックスは、必要に応じて形態素テーブルの中に記述される。通常の辞書における見出しの設定は、表記より上のレベルの音声上，意味上の一体性に基づいて行なわれる。この様な辞書は、語の構文的な特徴、意味的な特徴を記述するには効率的であるが、個々の表記における問題点等を記述するには適していない。これが、我々が形態素テーブルを設け、形態素認定の過程と語認定(従来の品詞的認定)の過程をわけた理由である。

(2)　形態素をその特徴に従って細かく類別したこと．
我々のシステムでは、形態素解析,構文解析の精度をあげるためと校正支援のための参考情報にするために、形態素を細かく分類している。一般の辞書で、同一の品詞名を与えられている語であっても、その接続上・構文上の性質を細かくみると明確な違いが見られることがある。たとえば、「本」,「素手」,「程度」は一般の辞書では単に名詞になっているが、こまかく見ると、「本」はさまざまな連体修飾を受けるが「素手」は受けない、「素手」に付く格助詞は「で」が普通である、「程度」は接尾語の様に使われるという性質上の違いが見られる。こういった性質を接続チェックや係り受けチェックに用いれば、文解析におけるあいまい性を減らすことができる。我々は、形態素の類別を四桁以内の英数字で表わし、最初の一桁が大分類を表わし、桁が下がる毎により細かい分類を与えるようにしている。これにより、接続チェック、係り受けチェック等を細かくも粗くも行え、効率的である。

(3)　接続チェックのための補助的な情報を各形態素に関して記述できるようにしたこと．
接続上の制約は、通常形態素の類に関して表現されるが、ある類の中の特定の語がその類の別の語にはない特殊な性質を持つことがある。たとえば、通常動詞は

Recognition of Morphemes in the Proof Reading Support System
Fumon Shimizu,Noriko Yakuden,Youichi Hidano,Teiichi Kashiwagi
Japan Electronic Dictionary Research Institute,Ltd.

形容詞の語幹に接続しないが、"過ぎる"だけは形容詞の語幹に問題なく接続する。この場合、解析ルールを一般化すると形態素解析の精度を下げるので、形態素テーブルの"過ぎ"の項目に、局所的に形容詞の語幹と接続可能であることが表現される。また、校正支援システムは、はなはだしく特異なひらがな書き表記にも、対処できることが望ましい。しかし、どんな言葉もひらがな書きされることがあるとすると、現実の問題として、例えば"かい"という表記は、"買い","飼い","回","解","会","階","貝","下位","甲斐","海"のどの言葉の表記であるか、それだけでは判断できないことになる。そこで、極端な表記は考えられるけれども、実際にでてくることはないという意味のフラグを形態素テーブルに付与できるようにしている。

3. 校正支援のための情報の提供

校正支援のための情報は、形態素テーブルの各エントリーに対して、を記述するようにしている。ここで、形態素テーブルで記述される情報のいくつか例を挙げる。

（1） 表記上の特徴の指摘

次の様な特徴が指摘される。

① 難解な漢字である。
② 常用漢字ではない。
③ 常用漢字の音訓としては、認められない。
④ 慣用的な宛字である。
⑤ 慣用的でない宛字である。
⑥ 漢字書きできる。意味による書き分けがない。
⑦ 漢字書きできる。意味による書き分けがある。
⑧ 漢字が誤りである。
⑨ 送り仮名が(類推的/恣意的に)省略されている。
⑩ 送り仮名が誤りである。

（2） 読みの特徴の指摘

次の様な特徴が指摘される。なお、[ ]の中の文字は、形態素とは、別物であるが、どの語の表記であるかを明らかにするために記した。

① 濁音化、促音化などが類推的になされている。
("がかり"等)
② 濁音化、促音化、長音化等が恣意的になされている。("あったか[い]"等)
③ 基本的和語のならびでの省略・読み変化である。
("じゃあ","て[る]"等)
④ 外来語の表記が標準的とは認めがたい。
("ヴァイオリン"等)

（3） 読み語構成の提供

すべての形態素に対して読みと語構成が示される。

例 どす黒－－－－－＞どす！ぐろ

（4） 標準的な表記の提供

① 標準的な漢字書き表記の提供

例 ほんだな－－－－＞本棚

② 難解な漢字や慣用的な宛字表記の提供

例 どこ－－－－－－＞何所

③ 正しい漢字表記の提供

例 収獲－－－－－－＞収穫

④ 正しい送り仮名の提供

例 美くし－－－－－＞美し

⑤ 標準的な表記の提供

例 て[る]－－－－－－＞てい[る]

例 じゃあ－－－－－－＞では

（5） 送り仮名、外来語表記に"ゆれ"がおこる可能性の指摘

送り仮名、外来語の表記に"ゆれ"がおこる可能性がある場合には、標準的な表記を一つ定め、各表記からその標準的な表記を参照できるようにしている。これにより、表記の"ゆれ"を容易に検出することができる。

例 表す
表わす
→ 表す

（6） 意味による漢字の使い分けに関して誤りのある可能性の指摘

意味による漢字の使い分けが難しい語は、漢字表記からひらがな表記を参照できるようにしている。漢字の使い分けの正しい判断は、当社で開発中の辞書を用いて行なう予定である。

例 合[う]
会[う]
遭[う]
遇[う]
逢[う]
→ あ[う]

4. まとめ

日本語文の校正支援においては、従来の形態素の綿密な解析と同時に、それとは異なる文字表記上の"ゆれ"についても綿密な解析が必要である。そこで、この目的に適した特種な形態素を考案した。本報告では必ずしも理想的に書かれていない日本語文に対して、上記した文字表記上の形態素を用いて表記レベルの解析を精細に行なう方法について述べた。

参考文献

1）清水他 「意味情報を用いた校正支援」,情報処理学会第３７回全国大会,６Ｂ－７(1988)
2）樺島 忠夫 「日本の文字」,岩波新書７５(1979)