3Q−2

文章校正ルールの適用率向上に関する考察

墨康成　柴田昌宏

(株)沖テクノシステムズ ラボラトリ

1　はじめに

現在，我々は文章校正支援システム*（以下，本システムと呼ぶ.）が持つ文章校正ルール（以下，ルールと呼ぶ.）の評価を行なっている．この評価は，ルールの適用率向上に有効な改良方法を見つけることを目的としている．

本稿では，まず簡単に本システムの特徴を述べ，ルールの適用率向上に関する調査方法と結果，追加した実験を示し，さらにルールの改良方法を述べる．

2　本システムの特徴

本システムの主な特徴を以下に示す．

- 文の解析手法に形態素解析を使っている．
- ルールに基づき文章の誤り箇所を指摘する．
- 誤りが規則的で訂正が容易なものについて，自動的な訂正機能を提供する．

本システムの構成を図1に示す．

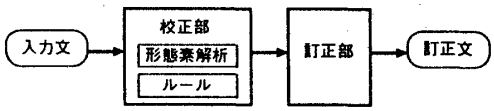

図1：システム構成図

3　調査の方法と結果

本システムにおける文章校正ルールの適用率を上げるため，以下の2つの方針を立てた．

- 出現頻度の高い誤りカテゴリに対応するルールに重点を置く．
- 適用率の低いルールのアルゴリズムを見直す．

これらの方針のもとに以下の調査をした．

[調査1] 誤りカテゴリごとの出現率に関する調査

仕様書，社内文書，電子メールなど，人間が見て誤りを含むと判断した文章203件を対象に調査した．これらの文章に含まれる誤りを25に分類し，さらに「文の体裁が悪い」，「文として成り立たない」，「読みづらい」，「意味的な誤り」と大分類した．そのカテゴリごとに誤りの数を数え上げ，誤り全体に対する割合を調査した．

調査した結果を図2に示す．

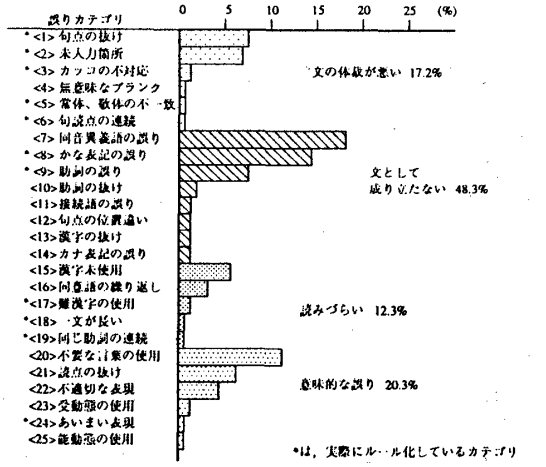

図2：誤りカテゴリごとの出現率

[調査2] ルールごとの適用率に関する調査

[調査1]で調査対象にした文章を本システムに与えて，誤りに対する各ルールの適用率（＝ルールを適用した誤り数／カテゴリごとの誤り数）を調査した．

調査した結果を図3に示す．

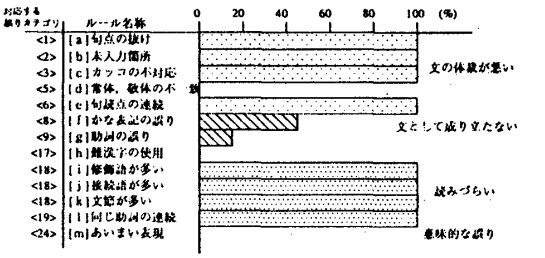

図3：ルールごとの適用率

[調査1に対する考察]

出現する誤りの約半数は「文として成り立たない」ものが占めていることが分かった．特に"<7>同音異義語の誤り"，"<8>かな表記の誤り"，"<9>助詞の誤り"が多く出現する．また「意味的な誤り」では，"<20>不要な言葉の使用"が誤りとして目立った．出現する誤りの例を図4に示す．

[調査2に対する考察]

「文として成り立たない」に属するルールの適用率が特に低く，改良を要することが分かった．(1)からも，この項目は出現する頻度の多い誤りであるため十分な適用率を得る必要がある．

3つのルール"[i]修飾語が多い"，"[j]接続語が多い"，"[k]文節が多い"は，どれも適用率が100%となっている．これらは，どれもルールを適用する基準値が低めに設定してあり，さほど長くない文に対してもルールを適用していることが考えられる．人によっても判断基準はあいまいであるため，人が読みづらいと判断する平均的な基準値を設定するほうがよい．

A Study of Elevating Application Rate of Proofreading Rule
Yasunari SUMI, Masahiro SHIBATA
Oki Technosystems Laboratory, Inc

*このシステムは，沖電気工業株式会社が開発したものである．

同音異義語の誤りの例
訂正が用意なもの → 訂正が容易なもの
かな表記の誤りの例
読みずらい → 読みづらい
不要な言葉の使用の例
不適当な部分(文が長いなど)などに → 不適当な部分(文が長いなど)に

図 4：出現する誤りの例

4 文の読みづらさに関する実験

人が読みづらいと判断する平均的な基準値を決めるために以下の実験をした．

[実験] 実際に社内で作成した論文スタイルの文章を被験者 15 人に流し読みしてもらい，直観的な理解度を

A. よくわからない
B. わかりづらい
C. わかりやすい
D. よくわかる

の 4 段階評価で採点してもらう．集計は，a) 文節数，b) 活用する自立語数，c) 接続語数 の 3 つの観点で行なう．

3 つの観点ごとに 4 段階評価の割合を計算してグラフ化した．それぞれの結果を図 5 ～ 7 に示す．

[考察] 仮に半数の人がわかりづらいと判断する点を境界値とすると，文節が 16 個以上，活用する自立語が 11 個以上ある場合にルールを適用するのが有効であると考えられる．接続語が増えると文は読みづらくなるようだが，わかりづらいと判断した人は半数以下だった．しかし，グラフの傾きから接続語の数が 4 個以上のときにルールを適用するのが有効であると推測できる．

5 改良方法

3. 調査の方法と結果 と 4. 文の読みづらさに関する実験 から，以下の改良が有効と思われる．

- **"[f] かな表記の誤り"** について
  現在のルール：ひらがなの未知語を検出するのみ．
  改良後のルール：さらに，誤りを含む文の人力による誤った形態素解析パターンをルールに付け加える．

例）漢字 一字の未知語 + ひらがな がある場合
読(未知語)/み(上一段動詞)/る(上一段動詞語尾)/よう/に/し/て/.

- **"[g] 助詞の誤り"** について
  現在のルール：未知語+助詞を検出するのみ．
  改良後のルール：さらに，誤った構文パターンをルールに付け加える．

例）連続する文節に同じ助詞がある場合
この/ファイル(普通名詞)/を(格助詞)//索引(普通名詞)/を(格助詞)//作成/する/.

- **文の読みづらさの基準値の変更**
  文節数の基準：10 文節以上 → 16 文節以上
  活用する自立語数の基準：4 個以上 → 11 個以上
  接続語数の基準：3 個以上→ 4 個以上

6 おわりに

今後の課題として，5. 改良方法 に示した方法でルールを改良し，その妥当性を評価することがあげられる．

謝辞
この研究にあたって集計に協力してくれた居倉裕二君と後藤誠君に感謝する．

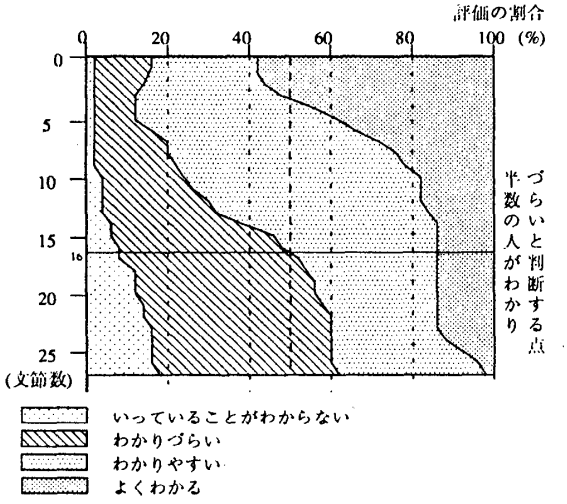

図 5：a) 文節数と文の読みづらさの関係

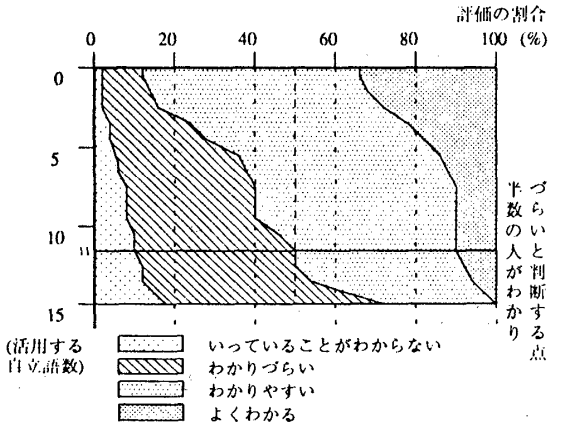

図 6：b) 活用する自立語数と文の読みづらさの関係

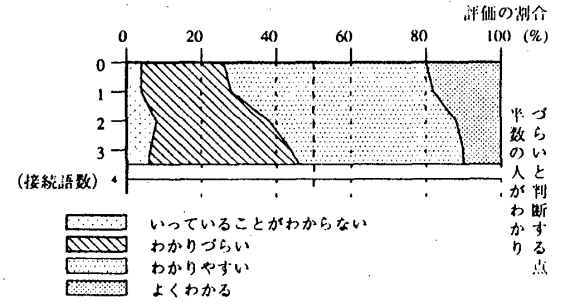

図 7：c) 接続語数と文の読みづらさの関係