Vol. 34 No. 10　　情　報　処　理　　Oct. 1993

解　説

IPS

自然言語処理技術の応用

2. 文書校正支援システムにおける自然言語処理†

池原　悟††　小原　永†††　高木伸一郎†††

1. はじめに

新聞，図書の出版分野などにおいて文書作成業務の計算機化が進む中，計算機化が困難とみられていた校正作業においても，ここ10年多くの研究開発[1)~13)]が行われるようになってきた．

日本文の校正は英語のスペルチェックと対比できる．英語が単語単位に分かち書きされ，単語の辞書照合が比較的容易に行えるのに対して，日本語はべた書きされるため，単語の辞書照合すら簡単ではない．誤字脱字などの誤りを精度良く発見するには日本語の解析技術の精度向上を待たなければならなかった．

日本語の校正支援の技術には，日本文を文字列データとみてデータ処理から接近する立場[3)]と，それを言語表現とみて自然言語処理から接近する立場[1),2)]がある．前者では，利用者の着目する観点からどんなデータをいかに抽出し表示するかが問題になるのに対して，後者では日本文解析の精度，すなわち，正しい日本文では，いかに解析を誤らず，誤った日本文の誤り部分を発見するかが問題となる．実用化されたシステムは，後者のタイプが多いが，誤りの検出漏れを防ぐために，日本文音声出力機能を組み合わせ，合成音声との対校方式を支援するもの[4),6)]もある．

一方，校正の対象とシステムの実現形態をみると，論文やマニュアルの校正[10)]を対象に，主にワープロの付加機能として開発されたシステム[4),7)~9)]と，新聞記事などの出版物用に開発され，専用のシステムとして実現されたものがある[6)]．

前者は一般利用者までの幅広い利用を狙った汎用システムであるが，当初はリソース上の制約もあり，一般の利用者の期待にこたえる機能の実装と品質の実現は困難であった．これに対して，利用者との共同開発が行われた後者のシステムでは，研究成果がフルに実装されるだけでなく，利用対象に依存した校正作業のノウハウなどが組み込まれることによって，実用性の高いシステムが実現され，校正現場で使用されてきた．最近では前者のタイプについても，ハードウェアの発展によって，実装上の問題は解決されつつあり，後者のシステムの経験をも取り入れたシステムが開発される[11)~13)]など，研究開発はますます盛んとなっている．

本稿では自然言語処理応用の立場から，文書校正支援システムが処理対象とする誤りの種類とそれに対する検出技術，訂正技術を取り上げ，実用化されたシステムを紹介する．また，今後の展望についても簡単に述べる．

2. 日本文誤りの種類

2.1 誤りの発生過程

通常，計算機内に文書ファイルが作成されるまでには，原稿作成，入力の二つの過程で誤りが混入する可能性があり，いずれも校正支援の対象となる．原稿作成段階の誤りは，あて字，差別単語などの文字／単語レベルの誤りから，文意の誤りに至るまで多種多様であり，その分布は，主に著者に依存する．一方，原稿入力段階で混入する誤りは，入力方式に強く依存する傾向がある．たとえば，新聞記事の入力で用いられるペンタッチ方式[14)]では，拾い間違いによる一文字単位の誤字脱字が多く[6)]，誤り方に文字配列の影響が認められる．これに対して，仮名漢字変換機能をもった

† Natural Language Processing for Japanese Text Revision Support System by Satoru IKEHARA (Knowledge System Laboratory, NTT Network Information Systems Laboratories), Ei OHARA and Shinichirou TAKAGI (Message System Laboratory, NTT Network Information Systems Laboratories).

†† NTT 情報通信網研究所知識処理研究部
††† NTT 情報通信網研究所メッセージシステム研究部

表-1 誤りの分類

分類	細分類		誤り例	特徴
【A】表記レベル	習慣性あり	【A—1】当て字／送り仮名誤り／カタカナ表記ゆれ	無理矢理→無理やり　行なう↔行う インターフェイス↔インタフェース	主として形態素解析処理により，誤りと断定できるもの
		【A—2】 俗語／禁止語	宅急便→宅配便　盲→目の見えない人	
		【A—3】 誤用語／誤用固有名詞	断圧→弾圧　廃除→排除 太田区→大田区	
	習慣性なし	【A—4】数値表記ミス／括弧などの非対応	一億4000万→一億四千万 (写真右→ (写真右)	
		【A—5】カタカナ表記ミス／字種変換ミス	オブジェット→オブジェクト チェックシテ→チェックして	
		【A—6】文法的に誤りとなる誤字／脱字	人間性→人間性 何を言っ販売したか→何を言って販売したか	
【B】表現レベル	単語(複合語)内	【B—1】 混ぜ書き語の送り仮名ミス	払込方法→払い込み方法 払い込み金→払込金	
		【B—2】 類型単語誤り	検索庁→検察庁　自造団体→自治団体	
		【B—3】 同音異義語誤り	処理公立→効率	
	単語，文節間	【B—4】 助詞誤り／脱落	東京でいく→東京へいく 計算機扱う→計算機を扱う	主として構文解析処理により，誤りと断定できるもの
		【B—5】 悪文[47]	犯罪を犯す→罪を犯す 改善する．～対処します．(不統一) ～しないと～しない (二重否定) 背の高い社長の椅子 (曖昧な修飾関係)	
【C】内容／一般常識レベル	【C—1】 実在しない固有名詞		中僧根元首相→中曽根元首相	文脈や一般常識を用いた解析処理により，誤りと断定できるもの
	【C—2】 矛盾する数値		第五四半期	
	【C—3】 文意に矛盾		物価上昇と運動→連動 定率法と低額法→定額法	
	【C—4】 文意の誤り			

ワープロ方式では，変換位置誤りや同音異義語誤りなどの単語単位の誤りが顕著となる[15),16)]．

2.2 誤りの分類

誤りはその内容からみて，おおよそ，以下の4通りに分けられる[17)]．

(1) 誤字・脱字や不統一な表記
(2) 不統一な表現と曖昧な表現
(3) 不適切な内容や構成
(4) 記述内容の過不足や不正確さ

新聞記事の場合は，通常，校正部門で(1)の問題を扱うほかは校閲部の問題とされる．学術論文などで査読委員が問題にするのはむしろ(3)，(4)の問題が中心である．

これに対して，文書校正支援システムの立場から，技術的背景を考慮して誤りを分類すると，おおよそ

[A] 表記レベルの誤り
[B] 表現レベルの誤り
[C] 内容レベルの誤り

の三つに分けることができる[18)]．それぞれの内容の例を**表-1**に示す．技術的には[A]，[B]はそれぞれ，形態素解析技術，構文意味解析技術の応用範囲と考えられるのに対して，[C]では対象知識や専門分野知識を背景とする意味理解以上の技術が必要とみられる．このため，現在のシステムが処理対象としている範囲は[A]と[B]の一部に限られる．

新聞記事などの出版物では，表記上の基準自体が非常に細かく規定されている[19)~22)]ためもあって，実際の校正作業は，専門的な知識と経験を必要とし，それに要する労力がきわめて大きい[11)]．そのため，[A]のみを対象としたシステムでも，大きな効果が期待される．

3. 誤り検出技術

3.1 形態素解析技術

日本語文書校正支援システムにおいて基本となる言語処理技術は形態素解析技術である．ただ

し，正しい文だけでなく，誤文の解析が問題になる．

（1） 正文を対象とした形態素解析

文を形態素（単語単位）に分割して，各単語の品詞などの統語的役割を決定する処理が形態素解析である．文法的に正しい日本文を対象とした形態素解析技術は，自然言語処理の最も基本的な技術として，早くから研究が行われ，機械翻訳[23]や文音声変換[24]などに応用されてきた．処理の手順は通常，以下のとおりである．

①仮文節（解析処理単位）の切出し

漢字自立部とかな付属部で文節が構成される場合が多いことに着目して，字種の変化点を手がかりに，仮の文節境界を決める．ただし，文節としては格助詞相当語「によって」なども付属部として扱った拡張文節[25]を用いる場合が多い．「は握する」などの漢字かな混じり単語は，後の処理で補正される．

②単語候補の辞書検索

仮文節内で単語候補の右方向最長一致法[26]などを用いて辞書と照合しながら抽出する．正解候補の抽出もれと不要な多義の排除とを調和させるため，単語の接続条件を考慮しながら，漢字列，仮名列部分の可能な単語候補を漏れなく検証する局所的総当たり法[27]などが使用されている．

③単語連鎖列を作成する

品詞などの接続条件が書かれた文法接続規則表を用いて，単語候補の列から正解とみなせる単語連鎖を作成する．漢字複合語の場合，文法接続規則表による制約だけでは多義があまり絞り込めないので，最長一致法[28]，文節数最小法[29]，DP 照合法[30),31]，単語間の意味的係り受け解析法[27]，単語共起要素法[32),33]など種々の多義解消法が併用されている．最近では構文解析結果から連鎖列を補正する方法[34]，確率モデルで多義解消精度を向上させる方法[35]などの報告がある．

（2） 誤りを含む日本文の形態素解析

形態素解析技術を校正支援に適用する場合，まず第一に，解析の正確さが問題となる．正しい文の解析を誤ると，後の確認訂正作業が増大し，役に立たなくなる．新聞記事の例では，正しい文の解析失敗箇所は，実際の誤り箇所の数以下であることが目安☆となる．

第二の問題は，誤りを誤りと検出できる精度の問題である．正文を対象とした形態素解析をそのまま使用した場合，以下の問題の生じることが指摘されている[36)~38)]．

（イ） 正文対象の文法接続規則表では，主に付属部で想定外の不当な単語列を正解と判定してしまう．例：「十分いなります」の場合，動詞「い（る）」と「なる」が接続不可と判定されないなど．

（ロ） 自立部で誤りを含む箇所を，短単位の不当な固有名詞や接辞として認定してしまう．例：「自浩(治)団体」の場合，「浩」が人名と認定されてしまうなど．

（ハ） 辞書未登録を原因とした未知語と，解析誤りによる未知語の発生が区別できない．例：「本だな(棚)」が未登録語の場合，「だな」は未知語と解釈されるなど．

これら問題に対し，校正支援システムでは，次の機能強化が図られる．

（a） 文法的接続検定機能の強化

付属語列で，前方および後方の二方向から文法接続判定を行うとともに，助詞の品詞カテゴリを細分化して，ひらがな文字の誤字／脱字が助詞候補として誤認定されないようにする．例：彼ををを訪ねた→「ををを」は接続不可．

（b） 漢字複合語に関する制約記述の充実

漢字複合語の単語列に誤字／脱字が含まれると，解釈で余った漢字が1文字の接辞や名詞に認定されやすい．これを避けるため，漢字複合語内の各単語の出現位置，単語間の意味的な係り受け属性などを単語辞書に記述して，この制約を利用する[36]．例：「肖像県（権）」において，行政区画「県」は意味的に「肖像」に接続不可と判定できるなど．

（c） 利用者辞書の導入と収録単語の強化

専門分野の用語などを登録できる利用者辞書を用意するとともに，常用外漢字のひらがな表記見出し（推敲→推こう），カタカナ表記見出し（米→コメ）などの単語を収録した辞書を用意する☆☆．

☆ たとえば，400字詰原稿用紙1枚（約200語）当たり2カ所の誤りがある文書の場合，形態素解析がその誤りを漏れなく発見したとしても，形態素解析の単語当たりの正解率が99%のときは，正しい部分で2カ所解析に失敗するから，校正担当者は合計4回確認訂正をしなければならない．さらに形態素解析の正解率が低くなれば，それだけ人手負担が増加し，むしろコスト高となる．

☆☆ この効果として，単語辞書の収録語数を8万語から50万語に拡大することで，文節分割誤りを1/7に削減できたとする例[39]がある．

3.2 表記レベルの誤り検出

(1) 習慣性のある誤りに関する検出

習慣性のある（再現性の高い）誤りでは，誤りやすい表記をあらかじめ辞書に登録しておき，それと対比して検出する方法が用いられる．その際，辞書には，誤りの表記と正しい表記（正規表現）を対にして登録しておけば，誤り発見後の訂正が容易である．

しかし，たとえば，検定規則として「日定→日程」や「断圧→弾圧」を辞書に登録した場合，この規則が「翌日定時配送」や「切断圧縮後」など，正しい表記にも適用されて，誤りと判定してしまう問題がある．この問題を避けるには，検定規則の適用条件を厳密にし，形態素解析と組み合わせた判定を行う必要がある．

検定規則の適用条件を厳密にするには，適用対象表現の文法的，意味的情報などの詳細な記述が求められる．しかし，一般の利用者が容易に登録できるようにするには，逆に，検定規則は単純な記述となるのが望ましい．幸い，このタイプの検定規則では右辺（書換え側）が正規の表現であり，システムの辞書から種々の情報が引き出せる．この点に着目して，利用者の入力した規則の適用条件を自動的に精密化する仕組みが実現されている[40]．

(2) 習慣性のない誤りの検出

習慣性のない誤りにも，まったく予想のつかないタイプと，ある程度予想のつくタイプの誤りがある．前者に対しては，あらかじめ登録しておくことが困難であるため，日本語解析技術に頼ることになる．すなわち，日本語解析プログラムが解析に失敗したところに誤りが存在すると仮定して，誤りを発見する方法[2]である．日本語解析プログラムとしては，表記レベルの誤りの検出では，**3.1** で述べたような，誤りに対して強化された形態素解析プログラムが使用される．

次に，ある程度予想のつく誤りでは，それぞれ個別に対策がとられる．以下に，その例を示す．

①ルールによる数詞表記の検査

数詞の表記では，算漢混合（一億4千万）や桁誤り（五万兆），位取り誤り（1,00），矛盾表記（1,2.3）などが比較的容易に検出できる[6]．その他括弧の非対応などの書式の誤りの判定も容易である[38),39)]．

②カタカナ表記ルールによる検査

カタカナ表記には，タイプミスなどで多様な誤りが出現する．この検査方法として，カタカナ表記を母音列と子音列に分解し（例：ヴァイオリン→子音列 VRN，母音列 AIOI），子音列どうし，母音列どうしで標準表記との類似性を比較する手法[41]，変形アルゴリズム（フェイ→フェー，ヴァ→バ，長音ー→削除など）で表記を同形化し，一致する原表記を誤りと検出する手法[39]などが実現されている．

③単語の文法的接続検査

単語が複合語を形成する場合に出現できる位置（先頭，中間，末尾）を単語辞書に登録しておき，不当な位置の単語を検出する[6]．

（例） 人問性→人間性

●「問」は複合語内の中間位置にこない．

文節末尾と次文節頭間においても，同様に文法的接続検査を行う．

（例） 何を言っ販売したのか

●促音便「っ」が誤り（助詞「て」の欠落）．

3.3 表現レベルの誤り検出

自然言語処理の応用の点からみると，表記レベルの誤りは，おおよそ形態素解析技術のレベルで処理でき，自動的に正誤判定のできるものが多い．これに対して，表現レベルの誤りでは，構文解析や一部意味解析までの技術が応用されるが，なお正誤判定は容易でない．

表現レベルの誤りは，表-1に示されるように，文節内に閉じて検出できるものと，文節間にまたがった解析の必要なものに分けられる．

(1) 文節内の誤り検出

文節内で誤りを検出する技術としては，以下の技術があげられる．

①混ぜ書き語の送り仮名チェック

新聞の表記では送り仮名は，後方にくる単語の品詞や名詞の意味属性に依存して規則化される[19),20)]．たとえば，同じ「はらいだし」という単語でも，後方に「抽象属性」の単語「方法」が接続する場合には「払い出し方法」となり，そうでない単語「金」が接続する場合には「払出金」となる．このように後方の名詞の性質を利用して送り仮名の検査を行う方法が実現されている[6]．

②複合語内単語の意味的接続検査

「検索庁→検察庁」のように複合語内で1文字

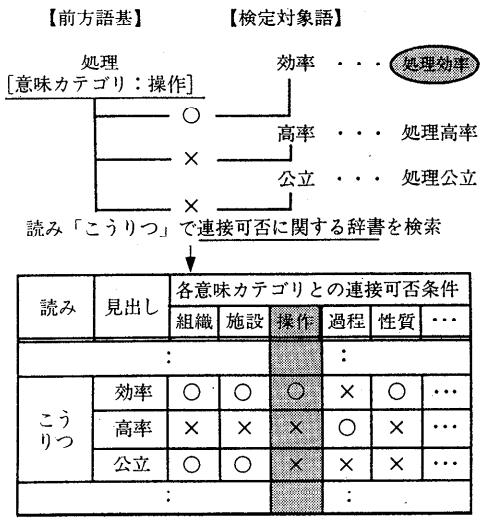


図-1 複合語内同音異義語誤りの検出技術

の固有名詞や一般名詞が係り受け関係を有せず存在する場合，誤字の可能性が大きい[36]．この特性を利用して誤り検査を行う．

③同音異義語の検査

同音異義語の字面とその前後に接続が可能な名詞の意味属性を記述した「連接判定テーブル」を使用して同音異義語の誤りを検出する手法が実現されている[42]．**図-1**に実現例を示す☆．

（2） 文節間の関係に着目した誤り検出

文節間の関係から誤りを発見するには，文節間の関係を解析する構文解析技術が基本となる．

①助詞の用法の誤り検出

助詞の用法誤りは構文解析によって発見するのが普通であるが，特定の助詞に限定し，その助詞固有の性質に着目して用法の誤りを発見する試み☆☆も行われている．

②悪文の検出

悪文と言われる文には，長すぎる文や埋込み構造が複雑で分かりにくい文[44),45)]などがある．この判定は，人間でも差のでる難しさがあり，現状で取り上げられている課題もその解決策も，まだ十分とは言えない．

以下では，REVISE—S[46)]に組み込まれている受動文の適切性判定の例を示す．受動文は能動文に比べて主張の強さなどに欠けることから，特にマニュアルなどの文章において好ましくないと言われる[47)]．しかし文脈的／統語的制約から，能動化すると解釈が変わってしまう場合がある．たとえば，「データが入力されていた」と「データを入力していた」を比べると，前者は完了の意味が強いのに対して，後者は継続の意味が強くなる．したがって，この場合，完了／継続のいずれかの意味を示す副詞がない限り，書換えはできない．このような観点から，態変換可否の条件を整理し，マニュアル校正に適用した例[48)]☆がある．

3.4 内容レベルの誤り検出

内容上の誤りを検出するには，一般常識や記述された対象に関する専門知識などを使用した知識処理の支援が必要となる．このレベルの誤り検出を網羅的に検出するシステムはまだないが，新聞記事校正の分野で，固有名詞の表す対象の実在性チェックを実現した例[49)]がある．

新聞記事には人名や企業名などの固有名詞が多出し，これが誤ったまま新聞が発行されると，訂正記事が必要となる．この例では実在する主な人名，地名，企業名などを辞書に登録し，形態素解析の未知語検出機能と組み合わせて，実在しない固有名詞を検出している．

3.5 音声合成による支援

前述の誤り検出技術のすべてを組み合わせても，現実の文書誤りを漏れなく完全に検出することは困難である[50)]．特に，表-1［C］の誤りの自動検出は難しい．

従来，新聞社などの人手校正では，二人一組みでの読合せ作業（対校）により，校正漏れを防止する工夫が行われてきた．この形態を校正支援システムに取り入れ，計算機が人に代わって読み上げた音声[51)]で校正作業を支援するシステムがある[4),6),40)]．

この例では入力された文書から，校正読みに合わせた音韻情報（音節）と韻律情報（ポーズ，アクセント）が生成され，合成音声装置を介して朗読音声が出力される．校正読みは，同音／類音語を読み分ける（「今秋」を「コンアキ」と読ませるなど），句読点，特殊記号などを読むなどの点で

☆ 図-1は，「操作」の意味カテゴリをもつ単語「処理」に接続できる単語「こうりつ」は「効率」でなければならないことを示している．この方法では，新聞記事中に出現する同音異義語127種の1,098件の誤りに対し，検出率86.5%が達成されている．

☆☆ たとえば，助詞「が」の場合，「直後に促音か撥音がくれば，それは助詞ではない．その直前が用言であれば，それは接続助詞であり，それ以外の「が」が格助詞である」という規則によって，接続助詞と格助詞が，それぞれ90%以上の精度で抽出された例[43)]がある．

☆ この例では，能動化が可能な文の76%，能動化が不可な文の87%が正しく判定されることが報告[43)]されている．

自然読みとは異なる．

なお，校正読みの支援においても，正しい文を誤って読まないことが，重要である．

4. 誤り訂正支援技術

検出した誤りには，自動訂正が可能なものと，自動訂正の困難なものがある．前者では，システムが誤り部分を訂正するが，後者では，極力，訂正候補を提示する．いずれの場合も，訂正誤りを防止するため，現状では，最終的に人手による確認が必要である．

4.1 表記レベルの誤り訂正

（1） 習慣性のある誤りに対する訂正候補の作成

3.2で述べたとおり，誤り表現と正規表現を対にして辞書[52]に登録しておき，該当する誤りが検出されたとき，正規表記に置き換える．この方法は，誤りの発見と同時に正解が得られるため，単純にして効果の大きい方式である．実用システムで多く使用されている．

（2） 習慣性のない誤りに対する訂正候補の作成

表-1 [A—4] の誤りは，書換えルールで訂正できる．それ以外の誤りについては，文法的に接続失敗した箇所で検出された未知語をキーとして，前後の単語の品詞などを手がかりに，以下のようにして，訂正候補が作成される[53]．

①カタカナ，英字用語辞書の照合

辞書から誤り単語の文字ごとに連想される単語を抽出して評価する連想統合型照合法[54]により類似度が高い単語を検索し，文法的あるいは意味的な接続関係を満たす候補を選択する．

例：プハジェクト→プロジェクト/サブジェクト/…

②文法的／意味的接続関係による訂正候補の作成

漢字二文字単語が多い点に着目し，一文字の漢字未知語からその漢字を含む二文字単語を抽出し，文法的あるいは意味的な接続関係を満たす候補を選択する[36]．表-1の [B—1], [B—2] などの表現レベルの誤りに対しても適用される．

例：検察する→検索/検討/…/考察…

この例では，検□/□察というサ変動詞型の名詞を選択するほか，動詞の語幹文字をキーに活用語辞書を検索して接続可能な送り仮名を提示する．

③マルコフ連鎖モデルによる訂正候補文字の抽出

大量の文を対象に，あらかじめ文字の連鎖確率を計算しておき，誤り文字の前後の数文字に着目して，確率的に接続の可能性の高い文字を訂正候補として提示する[55]．この方法は，かな文字列に比べて，漢字かな文字列で効果が大きいこと，また，1重マルコフモデル（2文字連鎖）に比べて2重マルコフモデル（3文字連鎖）の効果が大きいが，次数を上げても，漢字かな文での効果は頭打ちになることなどが知られている☆[56]．

4.2 表現レベルの誤り訂正

（1） 複合語内同音異義語誤りに対する訂正候補

訂正候補を，図-1と同様の意味属性関係を使用して検査し，条件が満たされる単語を選択する．

例：捜査性→操作性

（2） 文節間の誤りに対する訂正候補

文節間の文法的，意味的関係から発見された誤りは，誤りと言うより，むしろ不適切表現と言える．修飾関係の曖昧さを検出した場合に，利用者に修正案を提示する方法[57]も研究されているが，訂正は，必ずしも検出された部分に限定されないため，自動的に訂正するのは，かなり困難である．

この種の表現の訂正を校正と言うよりむしろ推敲の問題とみる立場から，文の改良と品質評価の試行錯誤を繰り返すことによって，質の良い文を生成しようとする試み[58]がある．しかし，現在の解析技術で文の評価を自動的に行うことは非常に困難である．

そこで，改良後の品質の評価は人に任せ，不適切と判断された場所ごとに，その理由と改良候補を示し，人手を支援する方法が実現されている[59]．この場合，修正項目の順序が計算コストに大きく影響を与えるため，この例では，形態素解析で処理できる誤りは構文解析に先だって訂正されるなど，負荷の軽い処理が優先して実施されている．たとえば，長文では，形態素解析の結果（連用中止，接続助詞など）から，複文や重文を構成する各単文相互の接続構造が決定され，適切

☆ 文字レベルでの連鎖確率に着目して正解を求める方法は，最近，隠れマルコフモデルとして音声認識候補の絞り込みなどで広く応用されている．

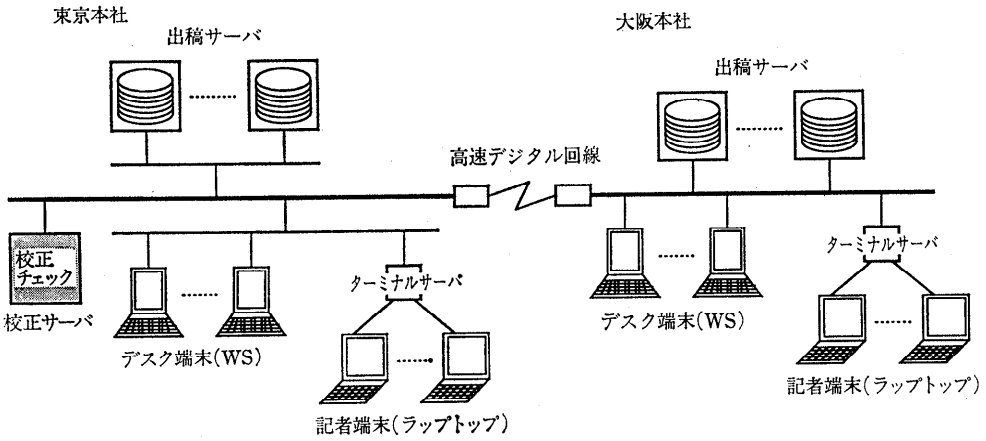


図-2 日本文校閲支援システム（第2期）のシステム構成

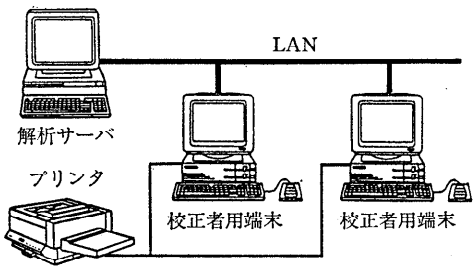


図-3 St. WORDS のシステム構成

な位置での分割，ならびに接続詞の補完が行われる[60]．この結果，長文は，構文解析処理に入る前に，適切な長さの文に分解され，後の処理の負荷が低下する．

5. システム事例

すでに多くの校正支援システムが実用化されている．ここでは，新聞，出版分野で使用されている3システムについて，その特徴を述べる．

5.1 日本文校閲支援システム (VOICE-TWIN)[13]

日本文校閲支援システムは REVISE[6]を母体としたシステムで，1987年5月より日経新聞社において稼働を開始し，一日 40〜70 万字にのぼる新聞記事原稿の校正に使用されてきた．その後，同音異義語の誤り検出と訂正候補の作成技術などの研究成果[37]を反映した第2期システムが1992年10月に北海道新聞社に導入され，日経新聞社においても1992年11月より試用を開始している．以下に，システムの特徴を示す．

①音声合成による校正読みの支援機能をもつ．

②新聞記事校正のノウハウが吸収されている．

このうち，①では文字当たりの読みの正解率99.8%が達成されており，②では，長年にわたり新聞社の校正部門に蓄積されてきたノウハウが，約7万項目にのぼる利用者辞書として整理され，さらに日々利用者の手によって更新されている．また，システム側では，利用者の校正用知識が容易に登録維持できるようにするため，**3.2** に示したシステム辞書情報による利用者辞書登録情報の自動補完機能などが実現されている．**図-2** は日経新聞社における第2期システムの構成である．

5.2 日本語文書校正支援システム (St. WORDS)[11]

St. WORDS は，COMET[4]を母体としてシステムで，**図-3** に示すように，誤り検出処理用の言語処理サーバと，文章の編集，管理を分担する校正用端末が LAN 接続されている．1992年6月末より，講談社の校閲システムとして試用されている．

特徴としては，広い分野の文書を校正できるようにするため，辞書登録語数が多い（50万語）こと[39]，口語的な表現にも対応できる[61]こと，形態素解析の処理速度が非常に速い（WS での処理速度350文字/秒[62]）ことが報告されている．

5.3 日本語校正支援システム (FleCS)[12]

FleCS は，1992 年 12 月より産経新聞社で実稼働を開始しているシステムで，パソコン単体上で動作する．本システムは，校正知識化する枠組みとして，校正パターン記述法と呼ばれる方法を用

いている．この方法では，校正担当者の校正知識を，比較的容易に計算機用のルールに置き換えることができると報告されている[63]．

6. 今後の課題と動向

新聞／出版業界を中心に，校正システムが現場導入されるようになってきた背景には，その基本となる形態素解析や構文解析などの言語処理技術が発展し，解析精度が向上してきたことがある．また同時に，利用者側の努力があり，校正部門の長年にわたる専門知識が集約されシステムに組み入れられてきたことの効果が大きい．

形態素解析や構文解析の技術で文書誤りを自動的に検出するには，きわめて高い解析品質が要求され，一般的な日本文に対して，このような技術を確立することは決して容易ではない．しかし，実用システムの経験からみれば，今後も適用分野を明確にし，その専門知識をも取り入れることで，基本技術と応用技術の双方の発展を図ることが期待される．

すでに述べたように，現在の技術で自動的に検出，訂正できる誤りは，まだかなり表層的なものに限られており，今後さらに，検出，訂正の精度の向上や処理対象誤りの範囲の拡大などが期待される．また同時に，専門業種との提携によって実現されたシステムのノウハウをより一般化し，一般の利用者を対象とした汎用的なシステムを実現することも大きな課題である．この点からみれば，LSI 化した形態素解析マシンをパソコンに接続する構想[64]が発表されるなど，ハードウェアの環境条件は一般利用者用システムに有利に発展している．

7. あ と が き

文書校正支援システムについて，主に実用化されている技術とシステムを中心に，その概要を紹介した．

紙面の都合上，単語辞書の内容と構成および維持・管理の問題，校正作業手順を支援するマンマシンインタフェースの問題，自動学習に関する問題については，あまり触れることができなかった．これらについては文献（インプリメント上の問題は，たとえば**5.** の関連文献）を参照していただきたい．

参 考 文 献

1) 池原，白井，神成：日本文の誤字に対する正解候補の抽出について，信学全大，No. 608 (1983).
2) 池原，白井：単語解析プログラムによる日本文誤字の自動検出，情報処理学会論文誌，Vol. 25, No. 2, pp. 298-305 (Feb. 1984).
3) 牛島，日並：日本語文章推敲支援ツール「推敲」のプロトタイピング，コンピュータソフトウェア，Vol. 3, No. 1, pp. 35-46 (1986).
4) 福島，大竹，大山，首藤：日本語文章作成支援システム COMET，信学技報，OS 86-21, pp. 15-22 (1986).
5) 池原，安田，島崎，高木：日本文訂正支援システム REVISE，情報処理学会第 33 回全国大会，4 J-9 (1986).
6) 池原，安田，島崎：日本文訂正支援システム (REVISE)，NTT 研究実用化報告，Vol. 36, No. 9, pp. 1159-1167 (1987).
7) 空閑：文書作成・校正支援システム WISE，信学技報，OS 86-28, pp. 13-18 (1986).
8) 鈴木，武田：日本語文書校正支援システムの設計と評価，情報処理学会論文誌，Vol. 30, No. 11, pp. 1402-1412 (Nov. 1989).
9) 高橋，吉田：計算機マニュアル推敲査読支援システム MAPLE の開発と運用，情報処理学会論文誌，Vol. 31, No. 7, pp. 1051-1062 (July 1990).
10) 大用昌之：次世代ワープロの決め手となるか校正支援/可読性評価ツール，日経バイト，No. 43, pp. 96-104 (1988).
11) 福島，佐々木，赤石沢，竹元：日本語文書校正支援システム，St. WORDS，情報処理学会第 45 回全国大会，6 C-1 (1992).
12) 奥村，脇田，金子：日本語校正支援システム「Flecs」，情報処理学会 NL 研究会，92-NL-87, pp. 83-90 (1992).
13) 北海道新聞社編集電子化推進本部事務局：道新ワープロネットシステム，新聞技術，No. 142, pp. 63-68 (1992).
14) 首藤，伊藤：ペンタッチ入力法，情報処理，Vol. 23, No. 6, pp. 536-542 (June 1982).
15) 中垣，山口，山下：ワードプロセッサで作成された日本語の誤りの分類の一考察，情報処理学会第43回全国大会，6 H-1 (1991).
16) 高木：ワープロ入力誤りにおける言語統計情報による誤り検出方式の検討，情報処理学会第 43 回全国大会，6 H-5 (1991).
17) 日本電信電話株式会社，他：マニュアル作成の技術，NTT ラーニングシステムズ株式会社 (1990).
18) 石井：新聞における校正・校閲の実データによる調査，ICOT Technical Report: TR-039 (1983).
19) 新聞用語懇談会編：新聞用語集，日本新聞協会 (1981).
20) 朝日新聞社用語幹事編：朝日新聞の用語の手引き，第 20 版，朝日新聞社 (1985).
21) 講談社校閲局編：日本語の正しい表記と用語の事典，講談社 (1985).
22) 共同通信社編：記者ハンドブック，共同通信社

(1985).
23) 長尾，他：科学技術庁機械翻訳プロジェクトの概要，情報処理，Vol. 26, No. 10, pp. 1203-1213 (Oct. 1985).
24) 宮崎，大山：日本文音声出力のための言語処理方式，情報処理学会論文誌，Vol. 27, No. 11, pp. 1053-1061 (Nov. 1986).
25) 首藤，吉村：日本語の構造とその解析，情報処理，Vol. 27, No. 8, pp. 947-954 (Aug. 1986).
26) 田中：自然言語解析の基礎，産業図書 (1989).
27) 宮崎：係り受け解析を用いた複合語の自動分割法，情報処理学会論文誌，Vol. 25, No. 6, pp. 970-979 (June. 1984).
28) 牧野，大澤：べた書き文の仮名漢字変換システムとその同音語処理，情報処理，Vol. 22, No. 1, pp. 59-67 (Jan. 1981).
29) 吉村，日高，吉田：文節数最小法を用いたべた書き日本語文の形態素解析，情報処理学会論文誌，Vol. 24, No. 1, pp. 40-46 (Jan. 1983).
30) 末田，金盛：動的計画法を用いた文字列と辞書項目の照合方式，信学全大，S 9-3 (1983).
31) 木暮，匂坂，嵯峨山，佐藤：日本語テキストからの音声変換における言語処理，信学全大，S 9-5 (1983).
32) 高橋，板橋：単語共起頻度を利用した形態素解析，情報処理学会 NL 研究会，88-NL-69, 69-5 (1988).
33) 大島，阿部，湯浦，武市：格文法による仮名漢字変換の多義解消，情報処理学会論文誌，Vol. 27, No. 7, pp 679-687 (July 1986).
34) 勘座：校正支援システムの日本語解析処理についての考察，情報処理学会第 46 回全国大会，3 L-1 (1993).
35) 久光，新田：条件付き確率最大法を利用した日本語形態素解析，情報処理学会第46回全国大会，1 B-1 (1993).
36) 安田，島崎，高木，池原：日本文訂正サービスにおける言語処理技術，情報処理学会シンポジウム「AI 技術の適用による新情報通信サービスの展望と課題」，pp. 79-88 (1987).
37) 小山，斎藤，小林，小山：文章校正支援機能における日本語解析，情報処理学会 NL 研究会，Vol. 69-2 (1988).
38) 清水，役田，披田野，柏木：校正支援における形態素の認定，情報処理学会第 38 回全国大会，5 E-2 (1988).
39) 福島，山田，小沢，他：校正支援システム St. WORDS の文書検査機能，情報処理学会第 46 回全国大会，3 L-3 (1993).
40) 合成音声で文章を読み上げる富士通の校正支援システム，日経データプロ・EDP 速報版 (1986. 7).
41) 小原，高木，林，武石：日本文推敲支援技術，NTT R & D, Vol. 40, No. 7, pp. 905-914 (1991).
42) 奥：意味カテゴリを用いた複合語の同音異義語誤り検定方式，情報処理学会第 38 回全国大会，2 J-7 (1989).
43) 下園，菅沼，牛島：日本語文章推敲支援ツール『推敲』における助詞「が」の抽出について，情報処理学会第 46 回全国大会，3 L-2 (1993).
44) 木下：理科系の作文技術，中央公論社 (1981).
45) 岩淵：悪文（第 3 版），日本評論社 (1984).
46) 林，高木：日本文推敲支援機能の実現方式，人工知能学会第 3 回全大，8-13 (1989).
47) 高橋：わかりやすいマニュアルの作成法，日経マグロウヒル (1985).
48) 林，千葉：日本語受動文の能動化可否判定アルゴリズムの検討，情報処理学会論文誌，Vol. 31, No. 10, pp. 1438-1443 (Oct. 1990).
49) 高木，安田，島崎，池原：日本語処理における固有名詞実在性検定方式の検討，情報処理学会全国大会 (1987).
50) 高木，島崎，池原：日本文校正支援システムにおける評価法の考察，情報処理学会第 37 回全国大会 (1989).
51) 宮崎，白井，大山，後藤，池原：日本文音声出力のための言語処理，情報処理自然言語処理シンポジウム (1983).
52) 島崎，安田，高木，池原：日本文訂正支援システム REVISE における辞書の構成法，情報処理学会第 34 回全国大会，6 X-3 (1987).
53) 高木，安田，島崎，松岡：日本文訂正支援システムにおける未知語訂正候補抽出方式，情報処理学会第 37 回全国大会，6 B-4 (1988).
54) 松尾，佐藤，津田：連想統合型照合による単語あいまい検索法，情報処理学会第 34 回全国大会，4 E-7 (1987).
55) 池原，白井：2 次 3 次混合マルコフモデルによる日本文誤字訂正候補の抽出，情報処理学会全国大会 (1986).
56) 村上，荒木，池原：日本文音節入力に対して 2 重マルコフ連鎖モデルを用いた漢字かな混じり文節候補の抽出精度，信学論 D-Ⅱ, Vol. J 75-D-Ⅱ, No. 1, pp. 11-20 (1992).
57) 箱守，杉江，大西：日本語の修飾関係を評価する添削支援システムを実現するための基礎研究，情報処理学会論文誌，Vol. 33, No. 2, pp. 153-161 (Feb. 1992).
58) 乾，徳永，田中：文章生成における推敲の役割，情報処理学会 NL 研究会，91-NL-83, pp. 47-54 (1991).
59) 林，菊井：日本文推敲支援システムにおける書換え支援機能の実現方式，情報処理学会論文誌，Vol. 32, No. 8, pp. 962-970 (Aug. 1991).
60) 武石，林：接続構造解析に基づく日本語複文の分割，情報処理学会論文誌，Vol. 33, No. 5, pp. 652-663 (May 1992).
61) 竹元，福島：口語的表現を含む日本語文の形態素解析，情報処理学会第 46 回全国大会，1 B-2 (1993).
62) NEC と講談社がソフト開発，朝日新聞朝刊 (1992. 10. 7).
63) 脇田，奥村，金子：日本語校正支援システム FleCS，情報処理学会第 45 回全国大会，3 F-4 (1992).
64) 大山：マニュアル検査システム TECS/M—基本構想—，情報処理学会第 43 回全国大会，6 H-3(1991).

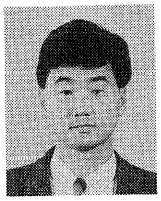
(平成 5 年 5 月 6 日受付)

池原　悟（正会員）

昭和19年生．昭和42年大阪大学基礎工学部電気工学科卒業．昭和44年同大学院修士課程修了．同年日本電信電話公社（現，NTT）入社．以来，電気通信研究所において数式処理，トラヒィック理論，自然言語処理の研究に従事．現在，NTT情報通信網研究所知識処理研究部主幹研究員．工学博士．昭和57年情報処理学会論文賞受賞．電子情報通信学会，人工知能学会各会員．

小原　永（正会員）

1954年生．1977年慶應義塾大学工学部電気工学科卒業．1979年同大学院修士課程修了．同年日本電信電話公社（現，NTT）入社，横須賀電気通信研究所勤務．以来，VLSIプロセッサの研究開発に従事の後，現在，NTT情報通信網研究所メッセージシステム研究部において日本文校正支援システム，日本文音声出力システムなどの自然言語処理の研究開発に従事．慶應義塾大学非常勤講師．電子情報通信学会，人工知能学会各会員．



高木伸一郎（正会員）

1956年生．1979年金沢大学工学部電気工学科卒業．1981年同大学院工学研究科電気工学専攻修士課程修了．同年日本電信電話公社（現，NTT）入社．以後，計算機アーキテクチャの研究を経て自然言語処理の応用をめざした研究実用化に従事．特に日本語の形態素解析技術をベースとした日本文校正支援システムの研究開発に携わる．現在，NTT情報通信網研究所研究企画部研究推進担当課長．電子情報通信学会，人工知能学会各会員．