デジタル・ドキュメント　5－2
（1997．1．31）

日本語文書校正支援ツールＨＳＰの開発

納富　一宏

神奈川工科大学工学部情報工学科

〒243-02 神奈川県厚木市下荻野 1030
TEL: 0462-41-1211　ex.3566
E-mail: notomi@ic.kanagawa-it.ac.jp
URL: http://www.ish.ic.kanagawa-it.ac.jp/~notomi

本稿では、解析速度に重点を置いた日本語解析支援ツールHSPについて、その基本アルゴリズムを簡潔に示した。特に、形態素解析に相当する処理を行なうために、文節、および文の表現モデルとして「JFKモデル」、および「JFKリンクモデル」を提案した。これらのモデルは、文字種別情報を利用して作成され、表記誤り、および統語的誤り検定のための前処理に適合するよう設計した。また、実際の動作例を示し、複数の文書を使った評価実験結果について述べた。評価実験から、HSPでは、毎秒1K[byte]程度の文書の誤り検出を行うことができる反面、平仮名列を含む部分の過剰検出が多く、解析精度に問題を有することが分かった。

Development of High Speed Proofreading Tools－HSP for Japanese Documents

Kazuhiro NOTOMI

**Department of Information and Computer Sciences,
Kanagawa Institute of Technology**
Phone: 0462-41-1211　ex.3566
E-mail: notomi@ic.kanagawa-it.ac.jp
URL: http://www.ish.ic.kanagawa-it.ac.jp/~notomi

In this article, I introduced a development of high-speed proofreading tools for Japanese documents (HSP), and proposed two models for error detection as a morphological analysis. One is JFK model for a Japanese clause -BUN-SETSU-, the other is JFK-Link model for a sentence. These models were designed to suit some types of proofreading. And also These methods were evaluated by 2 types of documents. It was confirmed that: HSP can perform a quick error-checking of about 1,000 [bytes] text per second on a smaller computer such as PC. But these methods have a risk of much detection of errors in HIRA-KANA strings.

1.はじめに

一般に、日本語文書の校正・推敲支援は、自然言語処理の初段にあたる形態素解析や構文解析の手法を用いることで、「ワードプロセッシング」を越えた知的支援を計算機によって実現しようとする狙いがある。

多くの場合は、対象となる文書に存在する表現の誤りを検出・訂正したり、より適切な他の表現に文章そのものを書き換えることを目的とする。

自然言語処理で問題となるのは、膨大な辞書データを必要とすることである。これらの辞書には、形態素解析に用いるものや、構文情報を持つもの、語の用例や慣用表現についてのテンプレートを持つものなど、多種多様であるが、これら解析に必要な多くの補助情報を効率よく利用しなければならない点が重要であるとされる。

個人レベルの文書作成では、数ページから十数ページ程度のものを処理できれば良いが、書籍やマニュアルなどの作成では、数百ページ以上のものをより効率よく解析しなければならない。

そこで、本稿では、パソコンなどの小規模システム上でも十分な解析パフォーマンスが得られる日本語文書校正支援ツールHSP(High-Speed Proofreading Tools)についてその解析手法を中心に述べる。

HSPは当初、高速性のみを追求したバッチ処理を行なうDOSプログラムとして開発された。現在では、インタフェース部分が大幅に改良され、よりインタラクティブなWindowsアプリとして稼働している。

以下、その動作例と性能評価の結果についても述べる。

2.データ表現モデル

本章では、HSPが利用するデータ表現モデルを示す。これは、日本語文を解析し、そこに現れた誤りを検出するためのデータ表現モデル（データ構造）であり、筆者らは「JFKモデル（JFK構造）」と呼んでいる。

2.1.文節表現モデル

JFKモデルでは、日本語における文節を表現することができる。文節には、自立部と付属部があり、形式的に句読点・記号部分を含む形態をなす。

JFKの各部分は、それぞれ表記上の差異を含む。すなわち、自立部（J部）は漢字、カタカナ、英数字からなる場合が多く、付属部（F部）は平仮名表記され、句読点記号部（K部）はそれ以外の非平仮名列からなる（図1参照）。

この意味は、JFK構造を得るために必要な情報は、被解析文の文字種別情報のみであるということである。

2.2.形態素表現によるモデル拡張

表記誤り、および統語的な誤り検出のために、JFKモデルでは、J, F, K各部をさらに細分化し、形態素レベルの表現ができるような拡張を施すことが必要である。すなわち、JFKモデルを形態素コンテナと捉えることができる。これを「拡張JFKモデル（構造）」と呼ぶ（図2参照）。

この拡張は、複合語解析および統語的接続検定の際に必要となる。また、この構造を得るためには、J部、F部ともにパターン辞書が必要となる。

2.3.文表現モデル

複数の拡張JFK構造を線形リストで連結して表現することで、文表現モデルが得られる。これを「JFKリンクモデル（構造）」と呼ぶ（図3参照）。

線形リストを用いるのは、文節という単位が、統語的には連結の自由度を持つからである。また、アルゴリズムを単純化するために、再帰を用いることが可能になるというメリ

ットがある。

2.4.文解析アルゴリズム

HSP における文解析アルゴリズムは、入力された文字列を拡張 JFK のリンクリストとして表現されるように、それぞれの部位を認識していくことを主体とする。このアルゴリズムの特徴は、文解析については、特に統語的な判定を含まないという点にある。唯一、文字種別（漢字、平仮名、カタカナ、英数字、句読点、カッコ、その他の記号）について分類を置き、平仮名列を挟む他の２つの部位を検出することで、JFK を抽出する。

従来型の解析手法では、①解析初段に統語的判定を含む、②解析に対して大量のデータを用いなければならない、という２つの理由のため、実行速度に問題を擁する場合が多い。

自立部判定や付属部判定において、10 万語程度の形態素辞書（「単語」よりも小さな単位である「形態素」が見出し語となる辞書。漢字、平仮名、カタカナ、英数字などすべての表記を持たなければならない）を用いるか否かで、処理パフォーマンスに差ができる。

当然ながら、従来型の形態素解析を主体とするアルゴリズムでは、解析速度を犠牲にする代わりに 9 割以上の精度向上が期待できる。しかしながら、標準的な日本語表現では、JFK 検出のアルゴリズムだけで 8 割以上の正しい解析結果が得られる。この差異がトレードオフとなる。

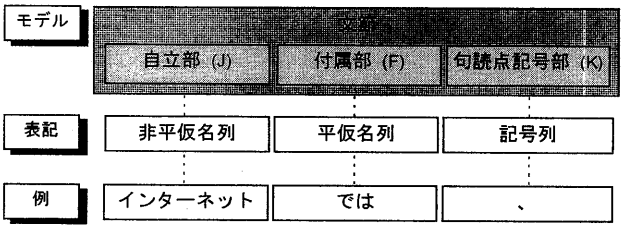


図１．JFK モデル

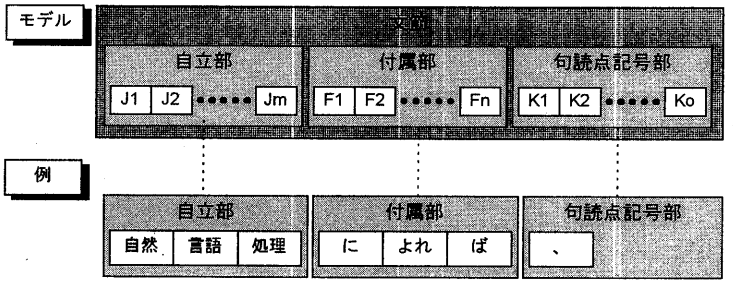


図２．拡張 JFK モデル

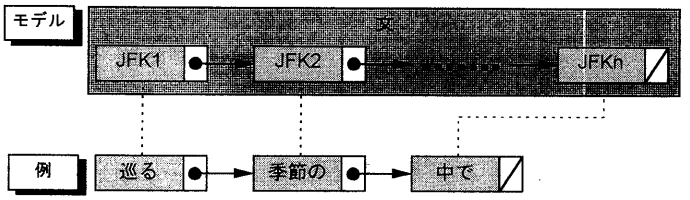


図３．JFK リンクモデル

3.表記検定・統語的検定

日本語の文章の誤り検出では、①表記レベル、②統語レベル、③意味レベル、④文脈レベルなどの別があり、この順番に解析が困難になる。実用レベルのものでは、①および②のものがいくつか存在する。

本章では、表記検定、および統語的検定のための解析手法について述べる。

3.1.表記検定

表記検定とは、一般に自立語を対象とし、あらかじめ用意された語群との照合を行うことで、リストから外れたものを検出することを意味する。表記検定は、MS-Word や一太郎などに代表されるワープロソフトにも採用されている文章チェック機能のひとつである。

このアルゴリズムは、非常に単純であり、チェックしたいリストを如何に用意すべきかのみが問題となる。たとえば、常用漢字表を用意すれば、これから外れる表記を検出できる。また、同音異字語リストを用意すれば、これにマッチした場合に特定のメッセージを表示することが可能である。さらに、誤り易い語のリストを用いれば、本来の誤り検出如何によらず、文章作成者に注意を喚起することができる。マニュアル作成などのように、表記のゆれを排除したい場合などに有効であるといえる。

この表記検定を JFK モデルに当てはめて考えると、これは自立部（J 部）の検定を行うことに相当する。

通常、辞書との照合を行うためには、辞書を検索するためのキーを決定しなければならない。たとえば、「インターネットでは、毎年多くのサイトが....」という文章について、「インターネット」という部分が正しい表記であるか否かを判定するには、まず、入力された文字列から「インターネット」の部分だけを抽出し、これから辞書を検索しなければならない。

こうした辞書引きのための前処理は、JFK モデルを用いる場合は、JFK 抽出段階ですべてなされている。よって、辞書照合が必要になった時点で、すぐに照合が可能である。

3.2.表記検定時の問題点

表記検定を行う対象は、すでに述べたとおり、自立語部分である。ここで、使用する辞書のタイプ、対象分野、規模などによって検出結果に差ができる。すなわち、正しい語であるにもかかわらず照合のための辞書が被解析文と整合がとれないため、誤りであると判定されてしまうことがある。

これを解消するためには、JFK モデルの J 部について形態素レベルの判定が必要になる場合がある。たとえば、「自然言語処理系分野では、...」という表記は、「自然」，「言語」，「処理」，「系」，「分野」というそれぞれの単語に分割可能である。これらの部分分割が通常の解析辞書で可能であれば、これを正しい複合化表現として判定することは可能である。よって、HSP では、自立部位の表記検定では、部分分割を許すか否かを指定できるようにしている。

さらに補足すれば、これは統語的には「複合語」として知られる。日本語のような膠着性の言語では、格語尾（格マーカ）としての助詞や用言の活用語尾が省略されることで、品詞転成が起こり、新たな語ができ易い。この性質は、辞書の構築を困難にする。通常は、未知語の自動学習などにより対処する場合が多い。

3.3.統語的検定

日本語の統語的要素は、表記における「平仮名」部分に集中している。通常、自立語要素は漢字表記やカタカナ表記が中心であり、特別な場合でない限り平仮名表記されない。これに対して、現代口語文では、付属語要素

は必ず、平仮名表記される。

自立部位の品詞の種類は、「名詞」、および、「動詞、形容詞、形容動詞など活用語の語幹」、一部の「副詞、接続詞」に特定される。

これらの統語的接続条件は明確に規定され、しかも複雑ではないため、すべてを列挙しておくことが十分可能である。すなわち、プログラム中に作り込んで置くことができる。

これに対して、平仮名表記部分には、多くの付属語要素が含まれる。品詞種別でいうと、「助詞」、「助動詞」、「形式名詞」などである。

日本語の膠着性を考慮しても、自立部位には特に注意すべき点は見当たらない。しかし付属部位には、本来の形態素解析と同様の文解析手法がほとんどの場合要求される。

形態素として付属語要素を見た場合、パターンの数は、300～400 パターン程度である。これらのパターンは 1 パターンにつき最大で平仮名 4 文字程度であるため、データ容量としても問題にはならない。

そこで、仮名漢字変換の高速化手法と同様、パターンの接続性を判定するために、HSP では、「パターン接続行列」を用いている。

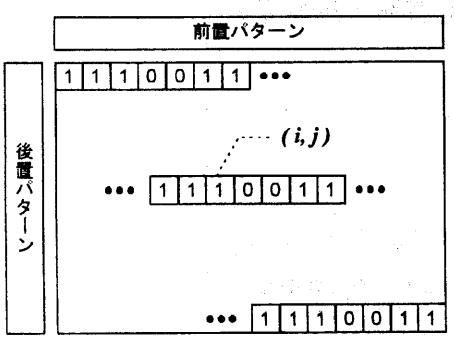


図４．パターン接続行列

これは、あるパターンと他のパターンとが接続可能であるか否かを２値表現したビット行列で表現したものであり、統語規則から得られた情報を基にあらかじめ作成しておく（図４参照）。実際には、422×422 のビット行列を用いている。

パターン接続行列を用いた付属部検定アルゴリズムは、概ね次のようになる（図５参照）。

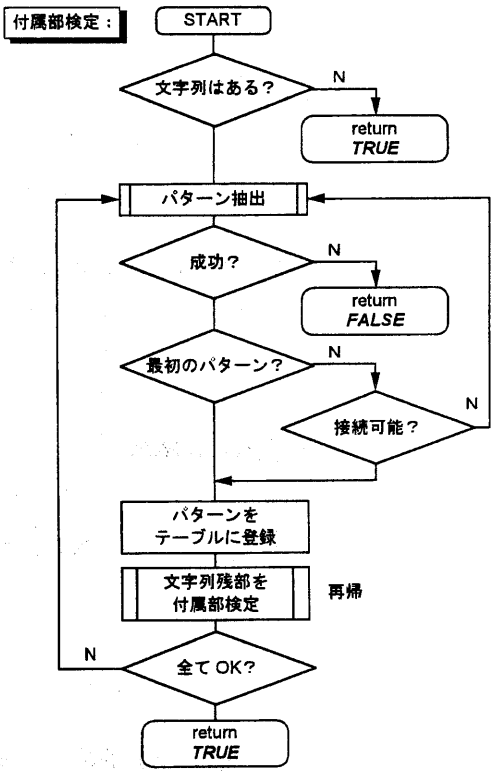


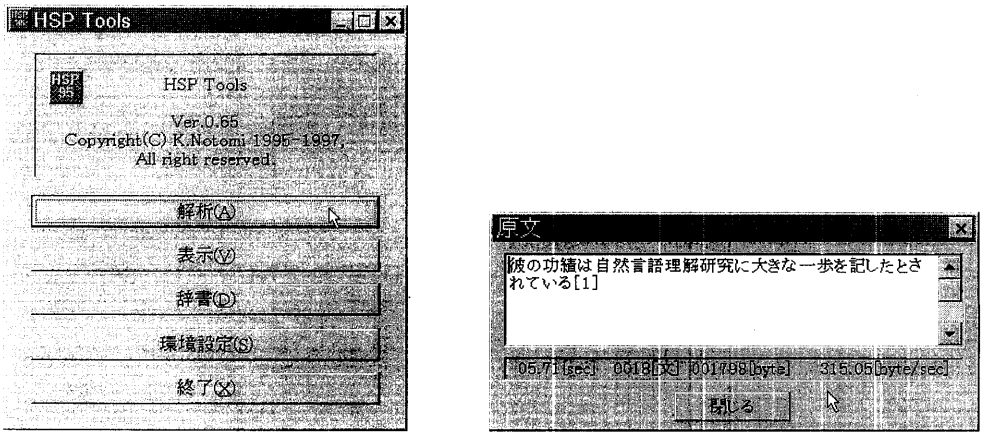
図５．付属部検定アルゴリズム

ここで、前置、および後置パターンの抽出は、最長一致法による切り出しを用いる。また、ループ表現部分を再帰表現に変えることも可能である。

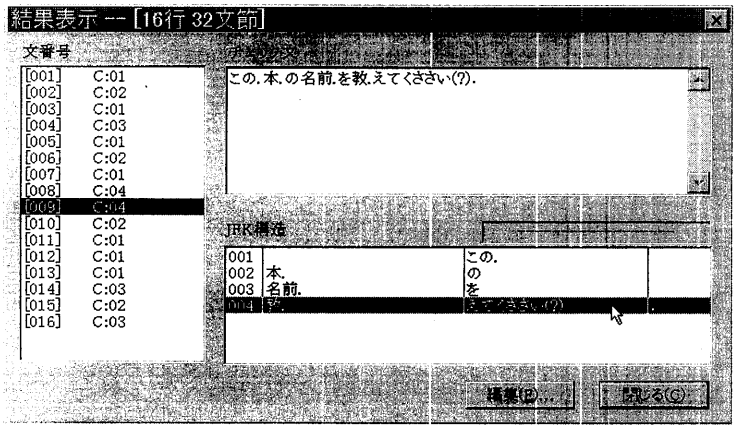
4.ツールの形態と動作例

HSP は、文書校正支援ツールとして、誤り検出機能を有する他に、解析辞書データのメンテナンス機能や文書処理に必要なフィルタツールを統合した形態をなす。

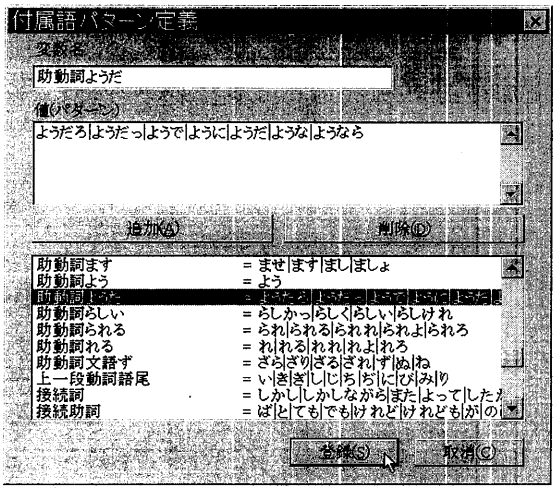
以下、動作画面例を示す。



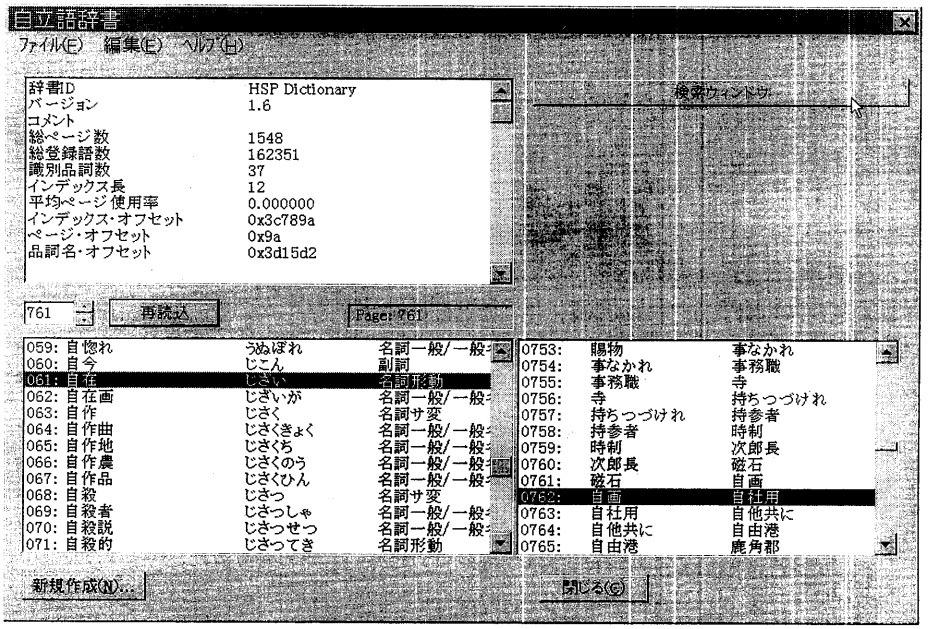
画面１．メニューダイアログ（左）および文章チェック実行中ダイアログ（右）



画面２．文章チェック結果表示



画面３．付属語辞書ツール



画面４．自立語辞書ツール

5.評価

本ツールの文章チェック機能を検証するために評価実験（CPU:Pentium-100MHz, RAM:40MB, PC/AT 互換機, OS:Windows95）を行なった。

評価に用いた文書は、文字データのみからなるプレーンテキストで、①自然言語処理に関する論文（na.doc）、および② Windows プログラミングのオンラインヘルプ（bc.txt）の２文書である。最初に、これらの文書から人手でエラーをすべて無くし、あらためてエラーを混入させた。エラーの種類は、①付属部への任意文字列の挿入、および②付属部からの部分削除である。また、自立部のエラー検出は解析に用いる自立語辞書の規模に依存するため、自立部へのエラー混入は避けた。今回の評価に使用した自立語辞書は登録語数 16 万 2 千語、付属語パターンは 422 である。

この条件の下、付属語解析を行なう場合と行なわない場合とで、解析速度および解析精度を調べた。結果を表１および表２に示す。

5.1.解析速度

複合語解析を行なう場合は、同じ単語について辞書へのアクセス回数が極端に増えるため、より多くの解析時間を要する。表１から、HSP の解析速度は、A4 用紙１頁を全角 1500 文字で換算すると、複合語解析を行う場合、および行わない場合では、１頁あたりそれぞれおよそ 3.1[秒]、および 2.0[秒]の解析時間がかかることになる。当然、文章表現の質に影響される。

5.2.解析精度

同じく、複合語解析を行なう場合と、行なわない場合とで、適合率、および再現率を求めた。但し、固有名詞等の自立語辞書における未登録語に起因するものはすべて除いてある。

適合率が極端に低くなっているのは、平仮名表記部分に関する過剰検出が原因である。HSP の行なう解析では、①自立語が平仮名表記された場合（例：りんご）、②平仮名表記

される接頭語・接尾語が存在する場合（例：お客さん）、および③活用語からの転成名詞が存在する場合（例：手続き）、にエラーの如何に関わらず、検出が行なわれてしまうという問題がある。再現率は、アルゴリズム上、複合語解析を行なう場合の方が低くなる。

適合率が低いということは、最終的な判断を人間がする場合、多くの箇所をチェックしなければならないということを意味する。これは、本来の計算機支援の目的とは逆の結果につながる。

特に、今回のようなインタラクティブなツールの場合は問題であろう。

個人レベルの小規模文書を扱う上では人間とシステムの双方向性は重要視される。しかし大規模文書の扱いをうまく行っていくためには、バッチ処理的な動作を行わせるか、あるいは、よりシステム自体が知的になる必要がある。

表１．解析速度

文書名	複合語解析	文数[文]	文書サイズ[byte]	解析時間[秒]	速度[byte/秒]
na.txt	○	216	21812	28.0	779.6
na.txt	×	216	21812	16.3	1338.2
bc.txt	○	37	3274	2.6	1273.9
bc.txt	×	37	3274	1.9	1714.1

表２．解析精度

文書名	総検出数	真の誤り	過剰検出	検出漏れ	適合率	再現率
na.txt	46	30	20	4	0.57	0.87
na.txt	52	30	24	2	0.54	0.93
bc.txt	14	10	5	1	0.64	0.90
bc.txt	20	10	10	0	0.50	1.00

6.おわりに

日本語文書構成支援ツール HSP について、その基本アルゴリズム、および動作画面例を示した。さらに評価実験結果について述べた。

HSP では、解析速度の点では問題がないと思われる。しかし、解析精度では、過剰検出の問題として平仮名表記語句への対処を行なっていく必要がある。

今後の目標は、現在、HSP とは別に開発を進めている複合語パーザ（順序依存による共起確率を用いた複合語検定プログラム、および複合語→平叙文プログラム）と格パーザ（格テンプレートマッチングプログラム）を HSP へ取り込むことである。

参考文献

[1] 長尾　真：“日本語情報処理”（第４章），電子通信学会，(1985-03).

[2] 片山朝雄：“三省堂実用 31 同音語選びの辞典”，三省堂，(1992.09).

[3] 納富，他：“自然言語処理を応用したマニュアル作成支援システム―マニュアル推敲支援について―”，情処学自然言語処理研報，(1991-09).

[4] 納富，他：“日本語文書校正支援ツールの開発―複合名詞の統語的検定について―”，情処学第 49 回全大論文集，3，3S-7，(1994.09).

[5] 納富，他：“日本語文書校正支援システムにおける高速統語解析手法”，神奈川工科大学研究報告，B 理工学編，第 20 号，(1996.04).

[6] Rodney G. Webster，中川正樹：“英語と日本語を対象にした文章誤り検出・訂正の共通点と相違”，情報処理，Vol.37，No.9，(1996.09).