# かな漢字変換辞書の製作
## ATOK2005 NHK新用字用語辞書

柴田　実

### 開発の目的

放送文化研究所用語表現班（当時）では2001年に『NHK新用字用語辞典第2版』を，さらに2004年に『NHK新用字用語辞典第3版』を出版した。（以下『用字用語辞典』と略記）

常用漢字制定後の，マスコミ界の変化を受けて，修正の必要が出てきたからである。

同時に，携帯用の電子辞書の普及や，各種辞書のCD版出版など，辞書をめぐる環境も大きく変わってきた。便利になったばかりではなく，逆に出版界の逆風もあり，本が売れなくなってきていることもその変化の1つである。

用字用語辞典は，NHK内部では「この漢字表記は使えるか使えないか」という場合や，「同音異義の使い分けは」と迷った場合などに多く利用されている。

掲載している語数（見出し語）はおよそ3万6千語であり，小型辞書である，『新明解国語辞典第6版』（三省堂），『岩波国語辞典』などと比べると語数は半分よりやや多い程度である。

これは，一般の国語辞典が固有名詞や外来語，複合語，動植物名，漢籍にある難しい語までを対象にしているためで，一般的な漢字表記に特化した用字用語辞典との性格の違いと説明できる。

現在，国語辞書や用字用語辞書を引きながら文章を「書く」場合に文字どおり筆記具で「書く」ことよりもワープロやパソコンを利用して「書く」ことが増えている。

ワープロ，パソコンで記述する場合，多くの人は意識するかしないかは別として，「かな漢字変換システム」というフロントプロセッサーを利用しているに違いない。

このかな漢字変換システムが正規表記に必要な変換を提供できれば，書籍の形での辞書に代わる道具として便利であると考えた。

このことには報道局システム開発も注目し，記者端末に搭載しているかな漢字変換をNHK仕様にしたいという要求が強かった。NHK仕様というのは，『用字用語辞典』に準拠した表記ができるということであり，放送でよく使われる表現や地名などが充実しているということである。

2003年に報道局システム開発の要請により，ジャストシステムのATOK16をNHK専用の辞書に変更したものを作成した。

これは，詳細な記述は避けるが，システム辞書を変更したため，ライセンスの問題で一般のパソコンでの使用には応用しにくい改変であった。

そこで，報道局限定版を開発した際に使用した，『NHK新用字用語辞典第3版』データを元に一般のウインドウズPCで使えるATOK市販版を開発することにしたわけである。

以下の記述にはかな漢字変換システムの説明があるが，辞書内容とともに，（株）ジャストシステムの公式な見解ではなく，筆者独自のものであるので本稿内容についてジャスト

システムへの問い合わせはご遠慮願いたい。

## かな漢字変換システムとは

1978年には東芝がワープロ専用機JW-10を発売しているが，ビジネス机くらいの大きさで630万円もしている。

日本語をコンピューターで使うのはあたりまえになっているが，個人向けのいわゆるパソコンが発売された1981年に三菱電機からMulti16という16ビットCPUパソコンが，日本電気（NEC）はPC8801という8ビットの機械が発売され，マニアのゲーム作りに使われていた。

その後，1982年に，本格的なビジネスユースをねらってNECからPC9801が売り出された。CPUはインテル8086同等品であった。16ビットマシンになってから，日本語をコード（数字）で扱うことが簡単にできるようになり，ディスプレーにも日本語が比較的容易に表示できるようになった。

しかし，基本的にはROM-BASIC（内部の固体メモリー依存）の機械であり，ワープロ機能も付いていなかった。

このため，プログラムを書く場合にも漢字一文字ずつを数字で入力しなければならず，非常に不便であったし，扱える漢字もJISの第一水準の6,600字あまりであった。

しかし，16ビットのCPUでもかなりのことができるようになり，単漢字入力から少しまとまった文章の変換が可能になった。

2年後の1983年には少し高価なPC100という機械が売り出され，これにはJX-WORDというワープロソフトが搭載されていた。

この時期が日本語かな漢字変換システムの草創期である。

漢字の扱いが容易になった背景には，漢字の標準的なコード体系化（JIS）とコンピューターのOS（オペレーティングシステム）がMS-DOSに事実上統一されたことがある。基準化と技術の進歩と言いかえてもよいだろう。

漢字の符号化を進めたのはJISのJIS C 6226-1978である。この制定にはさまざまな問題があったが，詳しくは他書に譲る。

JX-WORDのかな漢字変換はジャストシステムとアスキー（株）の共同開発によるKTISであるとされている。

漢字プリンターやディスプレーの問題を乗り越えながらの急速な開発であった。

開発は複数の会社個人で競うようにおこなわれ，1990年まで消長を繰り返す。

フリーソフトで公開されたXWP（エーアイソフト社）はパソコンユーザーに好意的に迎えられ，一方の商業製品は技術的に高いハードルを課されたが，それが全体の改善に大いに役立ったと言える（エーアイソフトは商業版のXWIIをのちに発売）。

1993年にジャストシステムは「一太郎6」というワープロソフトとともにATOK8（エートックはち・Advanced Technology Of Kana-kanji transfer）というかな漢字変換システムを発売した。そのATOK8は用例辞書を持ったAI変換[1)]を実現していた。

その後1987年にマイクロソフトからウィンドウズが発売され，1991年にウィンドウズ3.0となり，本格的なウィンドウズOS時代になる。

1995年のウィンドウズ95発売が日本での

パソコン普及の大きな節目になる。

オペレーティングシステムのMS-DOSからウィンドウズへのステップアップは厳しく，多くの社がかな漢字変換システム開発をあきらめ，大手だけが生き残った感があった。

同時にマイクロソフトがOSの元締めとしての立場を強固にし，ウィンドウズ用のかな漢字変換システムを製作するようになる。（IME・Input Method Editorと呼ばれる）

アメリカ資本による日本語システムの開発の時代に入ったわけである（ウィンドウズと異なったOSを持つマッキントッシュについてはこの稿ではふれない）。

2000年のウィンドウズ2000発売以後，かな漢字変換システムはマイクロソフトのMS-IMEとジャストシステムのATOK，エーアイソフトのWXG3，バックスのVJEなど限られた種類になっている。

ウィンドウズとともにバンドル（初期搭載）されているMS-IMEがシェアを誇っているが，その他の社のものも根強いユーザーがいる。

かな漢字変換の効率は，文章の解析と，辞書によると言われている。

解析はそれぞれの流儀とでもいうものがあり，その多くは企業秘密に属しており，公開されてはいない。

MS-IMEは辞書も公開されておらず，全体の表記法がどのようになっているかを検証することができないために，今回の開発の対象とはしなかった。

## 変換の手順

パソコンのかな漢字変換は，入力されたかな（直接のキーボード入力はローマ字であることが多いが）を「べた書き文」として扱い，文法を用いて単語に切り分ける。

切り分けられた単語を，助詞については中心部の変換機能で扱い，内部辞書を用いて動詞，形容詞，名詞などの変換候補を取り出し使用者に選択させる方法をとっている。

基本的な日本語文法を応用した形態素解析をおこなっているわけである。

辞書と打ち込まれた文字との比較については，さまざまな方法があり，「単漢字変換」（漢字1字ずつの読みによって変換するもので「買い物」と書きたい場合に「バイ」+「イ」+「ブツ」とひとつずつ変換するやり方），「単文節変換」（単語単位で変換するやり方），「連文節変換（複文節変換）」（助詞を含めた変換が可能），「全文一括変換」などがある。「連文節変換」「全文一括変換」では，多くの場合「二文節最長一致法」（入力されたべた書き文を辞書により文節に切った場合に，2つの連続した文節が最も長くなるような切り方になるようにする）という方法を用いているようである。

動詞，形容詞は活用形があるが，かな漢字変換辞書では語幹だけを持っており，活用の仕方などは別情報としてプログラムによって付加する形をとっている。この中心的な役割を果たすプログラム部分を「カーネル部」と呼んでいる。

そのほかに，結びつきやすい語の形を認識して変換効率を上げる「AI部」があり，「暑い夏に熱いお茶を飲む」のような文の場合，「夏」－「暑い」，「お茶」－「熱い」を結びつきやすくする変換を一部おこなっているが，この部分の辞書についてはあまり公開されていないし，使用者が効率的なユーザーAI辞書

を作ることは容易ではない。

今回，『用字用語辞典』を反映した，かな漢字変換を作るにあたっては，いくつかの点を中心に考えた。

1. 放送や，一般的な横書き文章に対してのかな漢字変換を提供するものである。

このため，ATOK 2005 オリジナルの標準辞書・トレンド辞書の登録語に対して，『用字用語辞典 』に準じた表記の正規化，かな漢字変換時に表示されるコメントの付与，登録語に対応する言いかえ語の設定等を施し，また用語の追加をおこなう。

2. その他の付属辞書（単漢字辞書，人名辞書，第三・第四水準漢字辞書）は特別な使用目的があるので正規化をおこなわない。

## かな漢字変換の仕組み

ここで，多くのかな漢字変換システムの働きを説明しておく。

かな漢字変換は，べた打ちかな入力の文章を文法的に解析し，単語に分けることから始まる。

そのためのプログラムがあり，さまざまな働きをしている。

活用のある語を語幹と活用語尾に分け，名詞と助詞を分離し，助数詞や接頭語などを判別する。

分けた単語を，辞書と照らし合わせ，照合がおこなえるまで文節に分けることを繰り返すこともあるようだ。

この部分は全くのブラックボックスで，企業秘密に属す部分である。

AI 処理といわれる用例比較や，特定の言い回しを判定するのもこの部分である。

このカーネル部が使用する辞書が「システム辞書」と呼ばれるもので，使用者（ユーザー）はなるべく手を加えないほうが効率がよいとされている。

これに対して，使用者側が特定の単語をたびたび使用したり，システム辞書にない単語を必要としたりする場合は，ユーザー辞書に登録して使用する。

プログラムであるカーネル部は，システム辞書とユーザー辞書をいわば 1 冊の辞書として扱うことができるようになっており，2 つの辞書を使い分けるわけではない。

同音異義語が多いのが日本語の特徴であるが，1 つの単語の読みに対して多い場合は数十の漢字が対応することがある。このような場合は，最後に使った変換候補を次のときに先頭に持ってきたり，使用頻度順に並べたりして，「変換ぐせ」を学習するのが普通である。

ユーザー辞書と性格が似ているが，使用者の判断でジャンルごとに大量の単語を登録した「専門辞書」を用意しているソフトもある。

医学用語や，経済用語，流行語，人名，地名，旧字体などさまざまな分野のものが市販あるいは，個人による提供などの形で流通している。

無制限に，何でも付加してしまうと，変換候補が多くなりすぎ非効率になるおそれがあるので，語の追加は使用者側の判断に任されている。

システム辞書については開発者の表記に対する考えや，カーネル部との整合性などで品詞の分類や登録漢字，送りがなについてのゆれがある。

WXII というソフトでは，送りの許容（省略や本則に比べて多く送る例など）を緩やかに考

え，その選択はユーザーに任せていたこともある。この場合は「おこなう」を変換すると「行う」という本則以外に「行なう」も候補に登場してきた。表記に対するさまざまな考えに対応することはできるが，逆に，使用者個人の表記の中にまちまちな表記が混在するおそれもあった。

最近は本則に統一されているが，名詞の送りがななどは複数の候補を含むものが多い。

システム辞書は世代を更新するたびにまちがいや本則以外のものは排除されてきたが，メーカーだけの努力ではなく，一般ユーザーの大きな声があったことは特筆に値する。

関係者の間では伝説ともなっている辞書のご意見番として，初期ユーザーに記憶されているのが箭内（やない）敏夫氏である。銀行に勤務するかたわら，自費でさまざまなかな漢字変換システムを購入し，製品テストをおこなった人物である。その結果は月刊パソコン誌に掲載され，後に，『電脳辞書の国語学』（おうふう）として出版された。

メーカーにとっては「日本語の標準表記が定まっていないのだから」という言い訳を許さず，手厳しい評価を下したことで，メーカーのまじめな努力を生む原動力の1つになったと言える。

## ATOK の辞書構造

ATOK の辞書構造はシステム辞書，ユーザー辞書とも基本的には類似している。

違うのは，システム辞書には単語解説とも言える注釈が2種類用意されているのに対し，ユーザー辞書では1種類だけであることだ。

2種類というのは変換した文字の後に赤い文字で注釈を付ける機能と，変換候補群ウインドウの横にさらに小さなウインドウを開き使い分けなどを説明する機能である。ユーザー辞書では前者しか実現できない。

ユーザーが扱える変換辞書は，次のような書式で記述する。

読み（ひらがな，カタカナの全角文字）

変換候補漢字（漢字，ひらがな，カタカナ，英数字の全角文字，英数字カタカナの半角文字，記号）

品詞（ATOK2005 の場合は 70 種類）

以上が必須項目

以下は任意項目で

コメント（変換中に単語について必要なコメントを標示させ，変換結果には含めない）

入力に対する置換をおこなうかどうかのフラグ

置換候補の 1～5

これらの項目を区切り符号（タブ記号）でつないだものを 1 行（1 データ）として記述したものが辞書のソース（元）になる。

変換辞書例

| | |
|---|---|
| 読み | # ひでただ |
| 漢字 | 徳川 秀忠 |
| 品詞 | 固有人他 |
| コメント | 贈り名は台徳院 |
| 強制変換の有無 | しない |
| 置換候補 1 | 台徳院（二代秀忠） |

「# ひでただ」と入力しスペースキーで変換すると，「徳川 秀忠」と変換候補が出てそのすぐあとに赤い字で「《贈り名は台徳院》」と表示され，変換候補群ウインドウには「→台徳院（二代秀忠）」と出てくる。

ユーザーは「徳川 秀忠」を選ぶなら次の語を入力するか，エンターキーを押せばよい。

「台徳院（二代秀忠）」を選ぶならもう一度

スペースキーを押して選択することになる。

（ここで「#」を付けたのは特殊な省略変換をするというユーザー側の任意の記号でありシステムとしては特に意味はない）[2)]

社員の姓名や，部署名，取引先の社名などよく出てくるものは省略形で登録しておけば入力の手間が省けるが，「か」だけで「加藤 太郎，笠井 次郎，金井 三郎，加勢 四郎」などが候補として出るのでは能率が悪く，「来年か」としたいのに「来年加藤 太郎」などという誤変換を引き起こしかねないので注意が必要である。

システム辞書に問題があるかどうか検証する必要があるが，ATOK2005 の場合は登録語数が約30万件あった。

これら 30万件を品詞別に見ることとした。品詞別の登録件数は**表1** にあげるようになっているが，品詞はいわゆる学校文法で習う品詞とは異なっている。

名詞を細分し，一般名詞と「する」が付くことがある名詞の「名詞サ変」，「する」のほかに形容動詞的な形があるものを「名サ形動」に分けている。実用的な文法に変更したと言える。

**表1**

| 品　詞 | 数 | % | 注　釈 |
|---|---|---|---|
| 名詞 | 110,934 | 37.56 | 一般名詞 |
| 固有地名 | 62,901 | 21.3 | 日本地名，外国地名 |
| 固有人名 | 28,286 | 9.58 | 人名のうち「名」の部分。外国は含まない。 |
| 名詞サ変 | 20,454 | 6.92 | +「する」の形をとる名詞 |
| 固有人姓 | 17,829 | 6.04 | 人名のうち「姓」の部分 |
| 一段動詞 | 16,231 | 5.5 | |
| 名詞形動 | 5,677 | 1.92 | 形容動詞形をとる名詞（圧倒的，アドバンス，荒れ放題，薄め，など） |
| 副詞 | 4,241 | 1.44 | |
| 固有組織 | 3,231 | 1.09 | 固有名詞の組織体名称（旭化成，宇宙研，営団，駿河屋，などの社名） |
| 単漢字 | 3,069 | 1.04 | 単漢字の読み |
| ラ行五段 | 3,060 | 1.04 | |
| サ行五段 | 2,825 | 0.96 | |
| 固有人他 | 1,932 | 0.65 | 人名で姓名に分けられないものや外国人名（阿倍仲麻呂，ケネディー，陶淵明，朝青龍など） |
| 接尾語 | 1,773 | 0.6 | 名詞に付く接尾辞（和え，祝い，請け合い，社間，省，島東端など） |
| 形容詞 | 1,664 | 0.56 | |
| 固有一般 | 1,640 | 0.56 | 一般的な固有名詞（嵐が丘，金沢城，紀勢線，クラリーノ，元亨，など） |
| 助数詞 | 1,265 | 0.43 | 助数詞（アンペア，Ω，円未満，回総会，日ごろ，シュート，ペニヒなど） |
| 独立語 | 1,257 | 0.43 | 単語だけで変換されるもの（相も変わらず，冠省，ガラーン，リットル，⊃（「けいさん」で変換），など） |
| ワ行五段 | 1,028 | 0.35 | |
| マ行五段 | 984 | 0.33 | |
| カ行五段 | 972 | 0.33 | |
| 形容動詞 | 899 | 0.3 | 形容動詞（あっぱれ，グローバル，知識的，面目躍如，ユニファイド） |
| 形容詞ウ | 400 | 0.14 | ウ音便をとる形容詞（呆気の，甘酸っぱ，七面倒くそ，雪深，など） |
| サ変動詞 | 398 | 0.13 | |
| 形動タリ | 347 | 0.12 | タリが付く形容動詞（唖然，春風駘蕩，断固，など） |
| 連体詞 | 339 | 0.11 | 有り得べき，ガリガリの，すべき，大それた，飲めや歌えの，など。「〜の」が多い（157） |
| 名サ形動 | 275 | 0.09 | 厚塗り，如何様，グニャグニャ，御謙遜，など名詞で「する」と「な」が付くもの |
| 数詞 | 261 | 0.09 | 幾千，一千，九百，13，四十，などの数字で形成される語 |
| タ行五段 | 240 | 0.08 | |
| 感動詞 | 179 | 0.06 | あーっ，お帰りなさい，グッドナイト，よっしゃ，などの挨拶語 |
| 接続詞 | 168 | 0.06 | |
| ガ行五段 | 163 | 0.06 | |
| 接頭語 | 138 | 0.05 | アンチ，押し掛け，高，被，などの接頭語 |
| バ行五段 | 109 | 0.04 | |
| 固有商品 | 55 | 0.02 | 味の素，クリープ，猫イラズ，マッキントッシュ，などの登録商標だが漏れも多い |
| ザ変動詞 | 51 | 0.02 | 甘ん，軽ん，そらん，嘆，など「ずる」が付く語 |
| カ変動詞 | 39 | 0.01 | 会いに来，見に来，やってこ，など「来る」の付く語 |
| 名詞ザ変 | 36 | 0.01 | 感，信，任，命，など「ずる」が付くもの |
| 冠数詞 | 17 | 0.01 | 金，計，午前，昭和，など次に数詞が来るもの |
| ナ行五段 | 8 | 0 | 溺れ死，凍え死，野垂れ死，など「死」の付く五段動詞 |

図１　ATOK2005 の品詞別単語数

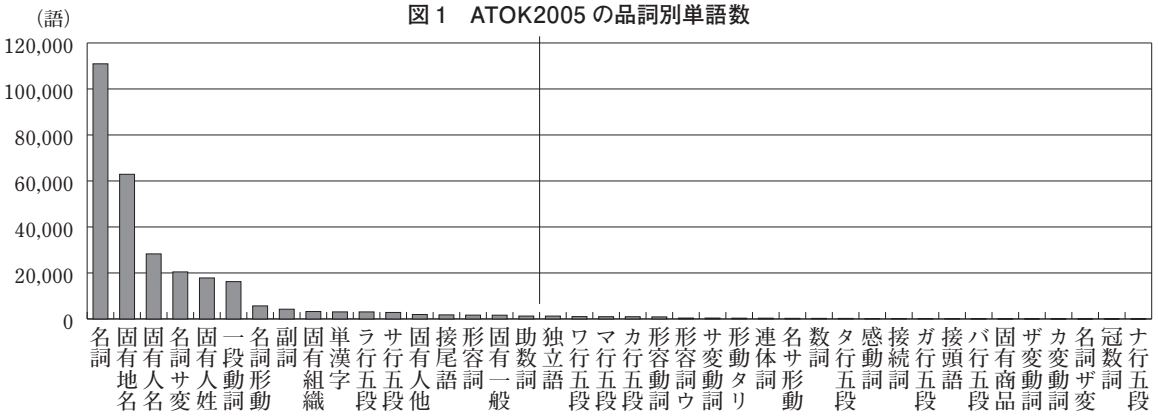
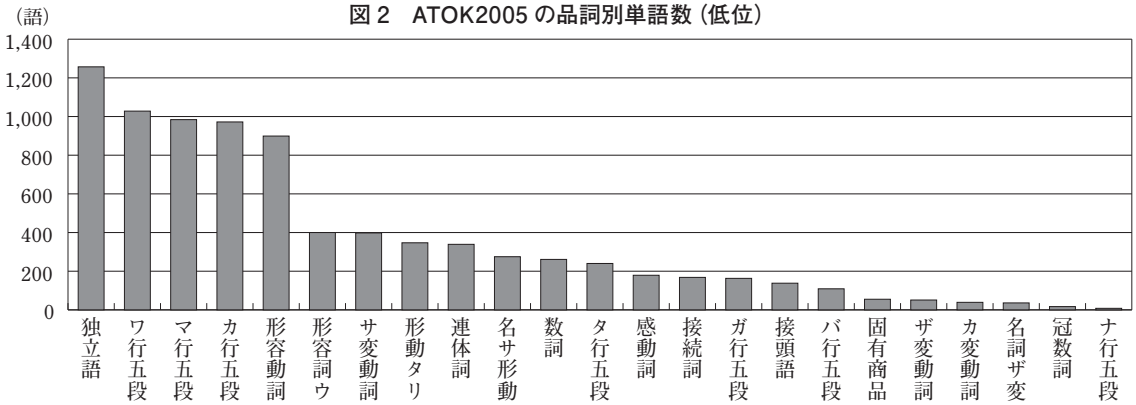



図２　ATOK2005 の品詞別単語数（低位）




固有名詞は７つに細分している。地名や人名，企業名などは後ろに付く敬称の違いや固有名詞の表記の違いがあるからだ。

活用のある語（用言）の分類は学校文法とほぼ同じであるが，例外的な活用をするものは極力少なくしてあるように見える。不規則活用は変換の中心部であるカーネル部分で処理をさせているために，少なくなっているように見えると推測できる。

動詞の活用が五段活用と一段活用に大別されているが，上下の一段活用を分けていないことも特徴的である。

さらに，文語の動詞で下二段活用があるが，現代文の中ではほとんど登場しないために，品詞としては用意されているが実際の登録はされていない。今後，文語用のワープロが登場するときのための準備と考えられる。

**図1，2** を見るとわかるように，

1. 名詞が圧倒的に多い（一般名詞と固有名詞で 86%）
2. 動詞以外の品詞はすべて 2% 以下の少数である（名詞形の多さにより比率としては低い）

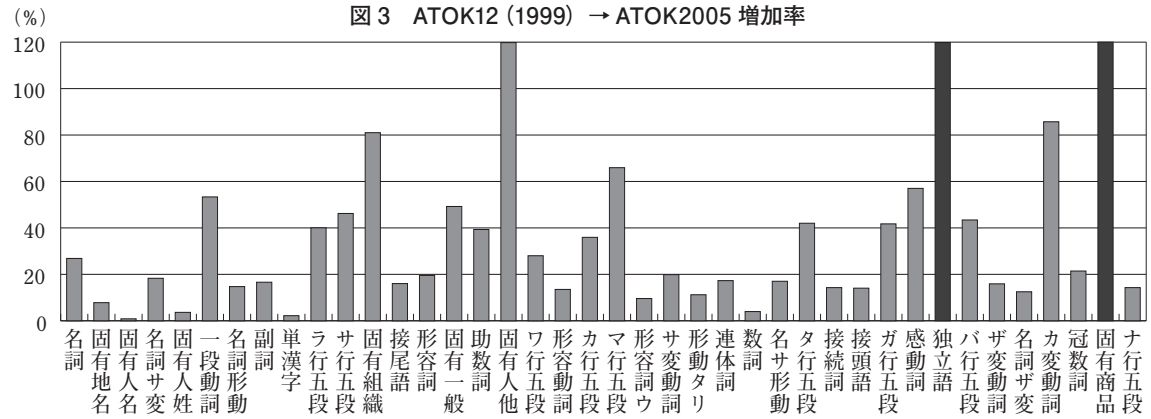
名詞系単語，付属語（接頭語，接尾語），単漢字をのぞくと３万５千語あまりである。

この部分が，日本語として文章の中で使われることが多い語群であると考えられる。

## どういう語が増えているか

ATOK12（1999）とそのあと６回の改訂を経

図3　ATOK12（1999）→ATOK2005 増加率




た2005とを比較すると，単純な増加数は47,514語であるが，割合として特に増えている品詞は独立語（84語から1257語と15倍ほどに増加），固有商品（12語から55語と4倍以上）がある。

独立語が増えたのは，四文字熟語の登録の増加と，口語的な言い回しの増加であろう。

固有商品はATOK12の母数が小さかったので，割合としては増えているが，単純増加（品詞変更を除く）では41語である。自社製品だけを登録していたものを拡大したための結果である。

カ変動詞の増加は「見に来る・会いに来る」など「来る」の複合語を増やしたためである。

そのほか増加率が高いのは固有人他（外国人名），固有組織（会社），感動詞などであり，一般名詞は20%程度増加している**（図3）**。

増加した語は，複合的な単語や，これまでに登録していなかった使用頻度の低い語であるので，基本的な部分の増加と言うよりは使い勝手の向上のための増加であろう。

名詞は20%の増加だが，実数では，1万9千語近くあり，これだけでも1つの辞書ができるくらいの分量である。

増加率が高い語種は，「固有人他」（外国人名，姓名が結合度の高い名前），「独立語」，「固有商品」でありいずれも倍以上増加している。

独立語，固有商品が急増しているのは，もともとの母数が小さいためと，変換実績から独立語扱いしたほうがよいと思われるものについて分類を再検討したものと見られる。

目立たないが，動詞の充実も集中的におこなわれていることが見て取れ，複合動詞，口語的な動詞が増加していることがわかる。

今後の方向としては，名詞や，固有名詞の追加が主なものになり，用言についてはさほど変更がないものと考えられる。

ATOK12から2005までのシステム辞書改訂を見ると，基幹部分の整備は終わり，追加をおこなっていると見ることができ，システム辞書はほぼ完成の域に達したと見ることができるだろう。

この点も，ATOK2005を専用辞書開発の対象にした理由である。

## 正書法への変換

システム辞書がほぼ完成の域に達したとはいえ，『用字用語辞典』に準拠しているかどうかはわからない。

そこで，システム辞書の点検をおこなった。

システム辞書を専用ツールを用いて，テキ

スト形式に変換し，データベース化して作業をおこなったが，提供されているツールではすべての語をテキストとして出力できるかどうかはわからないが，公式には「すべて出力される」ことになっているので，それに従った[3)]。

対象となるのは，固有名詞以外の語である。固有名詞はそれぞれの表記法があり点検は後回しでもよいからだ。

一般名詞が約11万語，その他が約7万語である。

名詞は数が多いのと表記の多様性があるためにチェックしにくいので，あとに回した。

まず単純に，動詞，形容詞などの用言を扱うこととした。

用言はATOK上では語幹が登録されていて，それに品詞情報で活用語尾を与える方法がとられている。一方の用字用語辞典では終止形が記載されている。

このため機械的な突き合わせができない。

そこで，活用語尾を品詞情報により補い，終止形を形成した。

この見出し語（漢字）データを，一語ずつ取り出し，さらに1文字ずつに分解し，それぞれの文字が，ひらがな，カタカナ，常用漢字，繰り返し符号，音引き記号などのどれかに含まれているかをチェックし，1文字でもそれ以外の文字（英・数字，半角文字，常用漢字表以外の漢字）であれば，データに表外フラグを付ける作業をおこなった。

詳細は省略するが，ビジュアルベーシックのプログラムを用い，対象表を検索することで実現できた。

次に，読みと漢字（見出し語）がATOKと用字用語辞典で一致するものを突き合わせ，データにチェックフラグを付与した。

チェックフラグがあるものは，一応，用字用語辞典に準拠している表記だと考えたわけである。

表外フラグがなく，チェックフラグがあるものは大丈夫と見られるが，全数に目を通し確認することができた。

用字用語辞典にある見出しがすべてATOKに搭載されているとしても，ATOKが名詞＋用言で18万語に対して，用字用語辞典は3万6千語であり，14万語あまりが対象外になった。

14万語については，語尾が同じものを集めたり，同じ漢字を使っているものを集めたりしながら検討を加えていった。

同様に名詞についてのチェックもおこなったが，『用字用語辞典』にはすべての名詞を掲載しているわけではないので，名詞の突き合わせ率は悪かった。

用字用語辞典が，一般的な名詞をすべて掲載しているわけではないことが主な理由である。

また，放送に登場するかどうかという観点がATOKにはないために，難しい単語表記が混じっていたり，明らかな文章語が含まれていたりして，判定に悩むことが多く時間がかかった。

## 許容の処理

用字用語辞典には「許容」という表記もある。

これは別にリストを作り，突き合わせ，備考に△マークを付与した。

たとえば「ことば」はひらがな書きを標準とし，「言葉」を許容としている。

報道局用のシステムでは標準形だけを認め，「言葉」は非許容という扱いにして，内部的な表記の統一を図ったが，一般に使用されることを考えると，「言葉」を非許容にすることは問題があると考えられた。

このため，変換時にはわずらわしいと思える注意コメントだが「言葉」には「△」印を付け，明示した。標準形を使う限りではこの注意コメントは表示されない。

また常用漢字表の付表には慣用的な読みを認めているものもある（小豆，海女，眼鏡，若人など114語）が，これらについては注意コメントを表示させず変換するようにした。

## 非正規表記の修正

正規表記ではない表記については，2とおりの修正方法がある。

かなを入力して変換キーを押すと，正規表記を強制的に表示し変換する方法（強制変換）と，表示は非正規表記であるがそれに「→正規表記」というコメントを付記し，そのまま確定すれば非正規表記に，もう一度変換キーを押すと次の候補として正規表記に変換できる方法がある。

後述する外来語の変換は強制変換を使用した場合があるが，一般的な漢字の変換についてはコメント付記・再変換方式をとった。

しかしこの方法では，同じ単語を何度変換しても「→正規表記」が出てきてしまう。

正規表記の登録がない場合には，かな漢字変換システムの「学習機能」（一度変換したものをシステムが覚えており，次の同じ単語が出てきたときには前に変換した候補を最優先する機能）が働かないからである。

このため，正規表記の単語がない場合には新たに登録する必要がある。

オリジナルのシステム辞書と，正規表記単語を対照し，ない場合には新たに単語登録をおこなった。

こうしておけば，変換候補の中から正規表記を選ぶと，次からは正規表記が最初の候補となり，スムーズに動くことになる。

この措置は，システムを知っている人には有効だが，単純に「→正規表記」のガイドに従って正規表記を選んだ場合には「→正規表記」が優先され，いつまでたっても正規表記への一回変換は実現しないので，使用上注意が必要である。

この現象を回避することを検討したが，難しいことばや不適切表現など別単語への言いかえが必要なものは，回避措置により本来の言いかえ機能が使えなくなるおそれがあり現時点では断念している。

## 具体的な作業

いささか詳細にわたるが，具体的な作業手順を紹介する。

まず，システム辞書の読み出しは，付属しているツールを使い，テキスト形式で書き出す。

このテキストを，マイクロソフトアクセスで扱えるファイルにインポートする。エクセルは，扱えるデータ量が65,500あまりで，ATOK2005の約30万件は限界を超えてしまうので使えない。

同様にして，用字用語辞典のデータも同一データベースにテーブルとして保存する。

アクセスには同じ項目を関連づけて2つ以上のテーブルからデータを抽出する機能や，VBAと呼ばれるベーシックプログラム言語も付属しているために言語データの扱いは格段にやりやすい。

一例として，見出し語に表外漢字を持つデータにチェックするプログラムを掲載する。

```
Function hyogai（）
Dim DB As Database
Dim RS, SS As Recordset
```

```
Dim DW, DD, CRET As String
Dim I As Integer
Set DB = CurrentDb
Set RS = DB.OpenRecordset（"Atok18", 2）
Set SS = DB.OpenRecordset(" 常用漢字 ", 2)
While Not RS.EOF
If InStr（RS! [ 品詞 ] , " 固有 " ）> 0 Then
GoTo ESC
If IsNull（RS! [ 漢字 ]）Then GoTo ESC
DW = RS! [ 漢字 ]
For I = 1 To Len（DW）
DD = Mid（DW, I, 1）
If Asc（DD）< 0 And Asc（DD）>
&H889E Then
CRET = " 文字 =" & " ' " & DD & " ' "
SS.FindFirst CRET
If SS.NoMatch Then
RS.Edit
RS! [ 表外 ] = True
RS.Update
GoTo ESC
End If
End If
Next I
ESC:
RS.MoveNext
Wend
End Function
```

以上のようにわずか30行ほどのプログラムで処理でき，30万件処理するのに10分以内で終了する。

プログラムによる処理を利用できるのは，統一した考えで検査するものであり，用法や場合により異なるというような不規則なものは適さない。また，送りがなの検査は見落としがあるので，いわば荒いフィルターにかけるようなものであることをわかっておく必要がある。

プログラムを利用して処理したのは

1. 表外漢字の摘出
2. 表外音訓の摘出
3. 送りがなのチェック
4. ニホン・ニッポン，地名の読み付け及びその読みとは別の読みによる登録

地名や，「日本」を含む名前の付いた企業名などの登録については，放送原稿で使用する場合に，アナウンサーが誤読しないようなくふうが必要であり，現場ではそのための地名辞典も用意している。この辞典を電子化できないかという要請もあり，次のような仕組みを考えた。

まず，その地名や企業名の読みを変換候補に付加することである。

この場合，何度でも出てくる場合には初めの変換時だけ読みを付加し，次からは読み無しにできないかを検討した。

これは，変換時のコメント処理をおこなうときに読み付きの変換を選択することにより実現でき，2回目からはそのコメントを無視すると（コメントのガイドに従わないことを意味する），読み無しの漢字だけの地名，企業名に変換できるので，使用者には1回目だけ選択をしてもらえばよいことになる。

例として，正しい入力「モノウチョウ」を入力した場合，変換候補は「桃生町≪読み付け（ちょう）≫」となり，もう一度変換キー（スペースキー）を押すと「桃生町（ちょう）」と変換できる。このときに変換キーを1回押しただけでは「桃生町」だけに変換できる。誤った読みによる入力（チョウとマチをまちがえた場合）は「モノウマチ」と入力すると変換候

補とコメントは，「桃生町《読み付け×マチ，〇チョウ》」となり，変換キーを2回押すと正しい読みが付いた「桃生町（ちょう）」となる。

全国には「マチ，チョウ」が入り交じった県も多く，なじみのない町名ではどちらが正しい読みかわからず，いちいち地名辞典を参照しなければならない。そのような手間を軽減するために，誤った読みでも正しく変換するシステムが実現できたと考えている。「ニホン，ニッポン」の読みも同じように処理した。「平成の大合併」による町村合併が多く，旧町村名を合併後の新地名に変換することも可能になった。

これらの処理のために，システム辞書には登録されていない「誤った読みによる辞書項目」を作成しなければならず，登録項目数の増加を招くことはやむをえないが，現在のパソコン能力を考えるとさほどの負担にはならないことからあえておこなった。

ただし，同じ漢字の町名が複数あり，県により「マチ」であったり「チョウ」であったりする場合がある。この場合は「マチ」で入力した場合は「≪チョウもあり≫」というコメントを付けた。また難読の市町村名には全体の読みを付けた。（例：「砺波市（となみし）」）

この機能は，システム自体から提供されるものであるが，辞書の作り方でさまざまな補助情報の提供が可能になる機能である。

## プログラム処理外の単語

プログラムで処理できないものについては，「未処理」フラグを設けこれらを一覧してチェックした。

この場合も細かな作業であるが，一部半自動プログラム（処理の途中で必要なつど人間の判断を取り込むプログラム）を書き，作業の補助とした。

このような手作業は6万項目あまりであった。

半自動作業で軽減できたのは，

1. 複合語後部要素（かける，つく など）
2. 常用漢字付表の熟字訓の複合例
3. 経済関係の複合語の特例

などであった。

複合名詞は後部要素をリストアップし，その後部要素を含むものをフィルターにかけ抽出し，検討を加えた。

常用漢字付表の熟字訓は「小豆，母屋，時雨」などの単語で，これらもフィルターにかけて抽出した。

経済関係の複合語の特例というのは，『用字用語辞典』の28ページにある「ウ．主として経済関係の語で，語尾に，「人，時，所，金，書，機関，制度，数量，品目」などを表す語の付くもの。

受取《人》売上《高》卸売《物価》貸越《金》貸出《金》貸付《金》借入《金》借越《金》 繰入《金》小売《商》差引《勘定》支払《人》 積立《金》取扱《所》取次《店》取引《所》※1)

引受《人》引換《券》※2) 振出《人》見積《書》売値 買値 問屋 仲買 歩合 両替 請負 裏書《人》 元売《価格》売手市場 買手市場 不渡手形 掛金 掛値 貸主 振込《金》

※1，2)「取引」「引換」は他の語の語尾に付く場合も送りがなを省いてよい。＜例＞商取引，代金引換

という部分である。

この部分は，網羅しているわけではないので別表のような表を作り，検討した。

37×138で5,106の組み合わせがあり，中にはありえない組み合わせもできるので，手作業であるものだけをチェックすることになった。

送りがなの許容とあわせて，このような特例の扱いは，日本語学習者にとってかなり負担になるであろう事がこの作業を通じてわかってきた。

## 正規表記の問題点

このようにして，固有名詞をのぞく辞書登録項目に対し，NHK ルールを適用してみたが，問題点もあらわになってきた。

1つは，表外漢字を含んだ熟語の扱いである。対応は4つのケースに分かれた。

1. すべてをひらがな書きにするもの。
2. 表外漢字をひらがなにして表内漢字は漢字のままとし，全体では交ぜ書きとするもの。
3. 表外漢字を用いるが読みを付加する（四文字熟語など）。
4. 表外漢字をそのまま用い，読みも付加しない。ただし，交ぜ書きまたはひらがな書きの変換候補を用意する（手紙文の結語や，伝統的に表外字を用いる慣用が特に強いもの）。

これらは，ケースバイケースで判断せざるをえないものが多かった。

交ぜ書きにする判断は，表内漢字を用いたほうが単語としての理解度が高くなるかどうかでおこなったが，常用漢字表に明確な基準がないためにひらがな書きと交ぜ書きの判断がゆれることがあった。これらの多くは，放送では言いかえをするか，用いない語であることが多かった。

常用漢字だけを用いた表記をしようとしても律しきれないものもある。

中国に語源がある四字熟語や，歴史の用語は交ぜ書きやひらがな書きはなじまない。

これらは，放送で使う場合はふりがなを付けたり，読みを解説することで理解を助けることをおこなっている。

今回の辞書製作では，3，4のケースはやむをえない措置であるとしたが，日本語表記の上ではさらに検討を加えなければならない問題である。

また，交ぜ書きやひらがな書きをすると思わぬ誤読を引き起こすものがあることも指摘しておきたい。「御」の扱いである。

「お，ご，み，おん」などの読みがあるが，接頭語としての「御」をかな書きにすると後部要素が1つの漢字であるようなものについては誤読を招きかねない。

「み心（御心）」が典型であり，「天皇のみ心」などと表記すると「天皇のみ＋心」と接頭語の「御」が前の単語の助詞とまちがえられるからである。

このような単語についてはかっこ内に読みを付けることで対処した。いわば緊急避難的な措置であり，多用すべきものではないと考えている。

今回のATOK2005辞書では単語そのものの言いかえも一部おこなっている。職業名の変更に伴うものや，省庁名の変更による言いかえなどである。

これを進めると，耳で聞いてわかりにくい単語の言いかえなどにも応用が可能であるが，どのレベル以上の単語を対象にするかなど課題もあり，現在の表記を用いるかどうかの判断はユーザーに任せている状態である。

正規表記の問題としては，数詞の問題がある。

かな漢字変換システムを使うのは横書きの場合だけではなく，縦書きの文書もある。

数を表す数詞は横書きでは算用数字を用いるが，縦書きでは漢数字を用いる。

縦書き，横書きでどちらの漢字を使用する

か自動的に判定することができず，今回は算用数字への変換を優先するようにしている。このため，縦書き愛好家には使いにくくなっている面は否定できない。今後の課題の1つである。

## 今後の課題

今後，ほぼ毎年内容のバージョンアップがおこなわれ，問題の解決も進むことが期待できるが，課題も残る。

1つは外来語の表記である。「チ」は「ティ」か「ジ」か「ディ」か，「ヴ」を認めるかなど未解決の問題が多い。在来の外来語と，新語の外来語では表記法がばらつくことはあり得るが，「慣用」としている表記も「原音に近く」という考えで別の表記を用いる人も少なくない（「ヘッドホン」と「ヘッドフォン」など）。

外来語，外国地名，人名の標準的なかな表記はどうあるべきかについては，増加する外来語への対応を含め問題である。

第2に，同音異義語の選択である。文脈から判断すればあり得ない同音異義語も機械的に変換候補として表示してしまう。「消化器」と「消火器」「小火器」は明らかに違う文脈で使用されることが多い。「消火器検診」などという誤変換を防げないものだろうか。変換候補選択の手間を省くくふうを考えたい。分野別の類義語辞典などを考えることが必要かもしれない。

第3に，ユーザー（放送）が用いている語彙はどのようなものであるか，固有名詞や数詞を含めた頻度調査が必要になるだろう。

通常使われる日本語の語彙範囲を知る上でも必要であり，多く使われる単語でカバーできる文章というものが，わかりやすい文章と重なることが予想されるからである。

この調査が可能になれば，個人の使用語彙も調査できることにつながり，日本語教育や，個人の文章力向上にも資することであろう。

今回試みたのはいわば「合理的な，電子書記法へのツールの提供」である。

多くの人が共通の表記を使用することで，データの検索や意思疎通が簡便になることにもつながると考えている。（しばた　みのる）

### 注

1）AI変換とは，人工知能変換と訳すことができる。単語1つずつを変換するほかに，単語が含まれた文章を解析し，より適切な変換を行うもので，用例をたくさん備えた辞書のような働きをする。文脈に合わない単語には変換しないようにすることができる。

2）入力の先頭に使える記号は限られているが，シフトキーを押さないで入力できる「@，¥」などが個人的な省略入力識別符号として用いられることが多い。

3）システム辞書には通常の方法では出力できない単語が存在することが推測される。「う」と「お」，「じ」と「ぢ」，「ず」と「づ」に見られる入力ゆれへの対処であると考えられ，2 − 3,000語の分量ではないかと考えられる。