◆特集 ……辞書をめぐる7つの闘い

6. 変換辞書をめぐる闘い

高本條治

ワープロソフトに利用される仮名漢字変換システムに備わる変換辞書は、普通の国語辞典とはずいぶんちがった仕組みをもっている。誤変換を少なくするために水面下で行われる変換辞書の「闘い」――その様々な工夫と問題点を紹介する。

一 仮名漢字変換システム

ワープロやパソコンで日本語入力をする際、現在最も普及しているのが仮名漢字変換と呼ばれる入力方式である。仮名漢字変換は、キーボード等から入力された「べた書き」の仮名文字列を解析して、日本語として妥当な漢字仮名交じり文に変換するサービスである。その際、仮名漢字変換サービスを行うシステムは、語の分割位置を決定し、個々の語の同定を行うことで、適切な表記形式をサービスしようとする。語の分割、語の同定、表記の付与のいずれかの段階で失敗すると、適切な変換結果が得られず、いわゆる「誤変換」となる。

仮名漢字変換システムは、基本的には次の三つの部分から構成されている。

・キー入力を受け付けたり、ローマ字仮名変換などの前処理を行ったりするユーザーインターフェース部

・べた書きの仮名文字列を解析・変換して、漢字仮名交じり文を返すプログラム本体部。「変換エンジン」と呼ばれる。

・変換エンジンが解析や検証に利用する変換辞書

仮名漢字変換システムは、利用者との相互作用的な協調作業によって変換サービスを行う。特に、日本語は同音語や同訓語が多いため、利用者は、システムが提示する複数の変換

候補群から自分がめざすものを選択したり、場合によっては、語の分割位置を区切り直したりしながら自分がめざす変換結果に接近していかなくてはならない。このような場合、ユーザーインターフェース部の使い勝手が、変換サービスの効率性に大きく関わる。

利用者が期待する変換要求に対して、どれだけ有効な変換結果をサービスできるかは、変換エンジンと変換辞書の総合成績で決まる。つまり、両者はいわば車の両輪として、変換サービスの有効性や効率性を左右するのである。いかに変換エンジンが優れていても、変換辞書がお粗末では話にならない。そのため、仮名漢字変換システムの評価では、変換辞書の内容や洗練度が常に問題にされる。それだけに、優れた変換辞書を制作しようとするとき、直面せざるをえない問題も多い。変換辞書制作の背後には、さまざまな「闘い」が見え隠れしている。

二　データベースとしての変換辞書

書籍版の辞書では、「国語辞典」「英和辞典」のように、正式な書名としては「辞典」を名乗るものが多い。それに対して、変換辞書はもっぱら「辞書」という名で呼ばれるのが普通であり、「辞典」と呼ばれることはまずない。この点はなかなか興味深い。

しかし、たとえ「辞書」という名称で呼ぶにしても、変換辞書を既存の書籍版辞典類との単純なアナロジーでとらえることには慎重であった方がいい。変換辞書は、あくまでも、変換エンジンが内部的に利用する機械可読のデータベースファイルであり、人間が直接参照するための可視的な文字で記された辞典とは、大きく性格が異なっている。

変換辞書のファイルは、慣習的にＤＩＣという拡張子（ファイルタイプを示す識別名）をもつことが多い。ＤＩＣはdictionaryの略であろう。変換辞書の実態は、読み方・表記形式・文法素性の三つ組を基本要素とするデータが集積された「語彙データベース」であり、付加的にそのデータに関連する情報や制約が記録されている。そのことを考え合わせると、ＤＩＣという拡張子は、Database・Information・Constraintの頭文字を組み合わせたものと考えてもいいのではないかと思う。

変換辞書というデータベースは、概念的には二次元の表であらわすことができるようなデータ構造となっている。一つ一つのデータ単位（レコード）は、基本的には「読み方」

「表記形式」「文法素性」という三つの属性（フィールド）から構成されている。この三つ組こそが、変換辞書というデータベースにおいて、最も重要なデータ単位を構成する。

一つのレコード内では、読み方と表記形式のペアは、その具体値が正規化されており、一つのフィールドが二つ以上の具体値をとることはない。こうして正規化された各レコードのことを、仮名漢字変換システムでは一般に「単語」と称している。ただし、「単語」の認定のしかたは、一般の国語辞典等に比べるとかなり異質である。例えば、活用語については語幹部分のみが「単語」の登録単位となるし、複数の語が連続した形態をあえて「単語」として登録する場合もある。また、読みと品詞が同じでも表記形式が異なっていれば別の「単語」として登録される点も、変換辞書の大きな特徴となっている。

三　変換エンジンによる辞書利用

変換エンジンは、ユーザーが入力したべた書きの仮名文字列を解析する際、解析中の文字列と変換辞書に登録されている読み方とを照合しながら、同時にその文法素性をチェックしていく。うまく条件に合致したレコードがあれは、表記形式を取り出して変換候補としてユーザーに提示する。大幅に単純化しているが、これが、変換エンジンによる仮名漢字変換サービスの流れである。変換辞書ファイルは、このような検索処理が高速に行えるように、巧みなインデックス化とサイズ圧縮の手法を用いてコンパクトに編成されている。

解析の際、変換エンジンは、活用語については活用形を派生させたり、また、接辞・助詞・助動詞・補助的文節などが連接した形式を派生させたりしながら、かなり複雑なパターンマッチングを試みる。そのため、変換辞書の文法素性には、品詞タイプの情報の他に、付属語等との連接パターンや活用語の活用パターンなどの細かい情報も含まれている。これらの情報は、変換エンジンによって参照され、パターンマッチングのための派生処理に利用される。

活用語尾のほか、補助的な文節となる語や付属語（助詞・助動詞）に関する情報は、変換エンジン内の派生処理アルゴリズムに組み込まれていることが多く、これらは変換辞書には登録されていないのが普通である。このような派生処理によって切り分けられる単位を、仮名漢字変換システムでは通常「文節」と呼んでいる。この「文節」も、学校文法でいう「文節」とは必ずしも一致しないところがある。

さらに、仮名漢字変換システムでは、「品詞」の概念やレパートリーも、学校文法などに比べると大きく拡張されている。例えば、ジャストシステムのATOK13には、次のような品詞が見られる。

・「安心」「邪魔」のように、名詞としても形容動詞語幹としてもサ変動詞語幹としても機能する語を「名サ形動」という品詞名で登録している。
・固有名詞は、「固有人姓」「固有人名」「固有地名」「固有組織」「固有商品」等に細分化されている。
・学校文法では区別されている「上一段」と「下一段」の活用タイプを一本化して、「一段動詞」という品詞に統合している。
・「接頭語」「接尾語」「単漢字」のように、通常は「品詞」とは認定されないものも、品詞として認定している。

品詞の概念やレパートリーの拡張は、変換エンジンが採用している特定の変換アルゴリズムによって要請されたものである。言い換えれば、どのような文法素性を変換辞書に記述するかということは、変換アルゴリズムの設計と表裏一体の関係にあるということである。

仮名漢字変換で最も重要な位置を占める形態解析処理には、「文節最長一致法」や「文節数最小法」といったロジックが採用されているが、どの手法を採るにせよ、不用意に変換辞書に登録された語が、別の語を変換させるときの障害となってしまうことがある。そこで、変換アルゴリズムからくる制約によって変換の障害となるような語は、「弊害語」として登録がしばしば見合わされることがある。

また、変換辞書に登録したままで何とか変換障害を回避するために、共起や接続のしかたを制限することもある。具体的には、いわゆる格フレーム情報に基づいた表層的な統語関係のチェックや、語と語の意味的な共起関係や接続関係のチェックが行われる。このとき、変換辞書には、別の語との共起関係や接続関係を条件づける「制約情報」が記述されている必要がある。検索されたデータレコードに、制約情報が定義されている場合には、その制約条件にしたがって変換候補の検証が行われ、例えば、第一候補を入れ替えたり、妥当性の低いものを候補から外したりするといった処理が行われる。

弊害語のチェックにせよ、制約情報の記述にせよ、このような辞書内容の細かなチューニング作業は、変換アルゴリズムの特性とふるまいをよく見通した者にしかできないし、ま

た、多大な時間と労力を要する地道な作業である。変換辞書における最大の「闘い」は、実は、変換アルゴリズムそのものとの間で繰り広げられている「内なる闘い」なのである。

四　「長単位」か「短単位」か

ごく単純に考えれば、変換辞書にはよりたくさんの語が登録されていた方が、便利であるように思える。仮名漢字変換システムが市場に広く出回り始めたころは、カタログや広告で「辞書登録語数」の多さを競う傾向があった。確かに、辞書登録語数の多さは、広範で多様な変換要求を充足させる上では重要な要因となる。しかし、変換サービスの効率性ということを考え合わせると、実際にはそれほど単純ではない。

辞書登録語数を大幅に拡大したとき、次のような問題点が出てくることが予想される。

(A)　新しい読みの単語が大量に追加されることによって、変換アルゴリズムが潜在的にもっている誤変換の可能性が顕在化してしまうかもしれない。つまり、随所で変換障害を起こす新たな「弊害語」を多数生み出してしまう恐れがある。

(B)　同じ読みの単語が追加されることによって、変換候補の数が増大し、正しい候補を選択するために利用者側が多くの注意力を払わなくてはならなくなる。つまり、利用者側の労力負担の増大を招くことで、効率性が低下する恐れがある。

(C)　辞書ファイルのサイズが巨大化することによって、それを格納するためのディスク資源が余計に必要になり、また、ファイルアクセスに要する時間も増えるので、変換速度に悪影響が出る。やはり、ここでも効率性が低下する恐れがある。

このように、辞書の登録語数を増やしさえすれば変換サービスの質が向上するというものではない。変換辞書と変換アルゴリズムは切っても切れない仲にあるので、辞書の力だけで機能的なブレイクスルーを起こすことはできないのである。

第二節で、変換辞書には、複数の語が連続した形態をあえて「単語」として登録する場合があると述べたが、このような単語を「長単位」の語と仮に呼ぶことにしよう。仮名漢字変換システムでは、変換サービスの有効性を高めるために長単位の語が意識的に登録されることがある。例えば、大阪市に「此花（このはな）」という区がある。「此花」が変換辞書に登録され

ているとき、変換アルゴリズムによっては、このままでは「この花」や「この鼻」が変換できない。その場合、「この花」や「この鼻」を長単位の語としてあえて変換辞書に登録しておくという措置がとられることがある。

また、変換サービスの効率性を高めるために長単位の語を登録することもある。「藤原道長」は「ふじわらのみちなが」というように「の」を入れて読まれるのが普通である。このような場合に、「藤原」を「ふじわらの」という読みで辞書登録するやり方も考えられるが、限られた人物名でしか「ふじわらの」という読みが使われないのであれば、「ふじわらのみちなが」という読みで「藤原道長」を登録するというような、長単位による個別処理の方が効率的である。

「此花区」という区名の場合も、「此花」の文法素性の中に、この語が区名である属性を定義しておいて、形態解析時に動的に派生させるという手法や、「此花」と「区」を共起制約で結びつけるという手法も可能だが、おそらく「此花区」を長単位の語として登録した方が効率的であろう。

一般の複合語や慣用句についても、長単位による辞書登録によって実現するか、それとも、短単位要素を組み合わせて動的派生や共起制約によって実現するかは、変換サービスの効率性を考慮した上で判断されているようである。

五　「規範性」か「記述性」か

仮名漢字変換システムの変換アルゴリズムとは別に、利用者の側にも、一人ひとりの利用者ごとに、言葉の運用基準という一種のアルゴリズムがある。しかも、同じ個人であっても、作成する文書のタイプや内容や目的によって、言葉の運用基準にもある程度の変動幅が認められるようである。

例えば、公用文書を作成する際のように、「常用漢字表」「現代仮名遣い」「公用文の書き方」等に完全に従った用字用語で文書を作成しなくてはならない場合がある。つまり、表記や表現のしかたに規範性が求められる場合である。その一方で、仲間内のメールのやりとりなどでは、規範に拘束されない表記をむしろ積極的に選択したり、くだけた口語的表現や方言的な言い回しを自由に使ったりしたい場合もある。

汎用の仮名漢字変換システムには、このどちらの場合にも一定水準以上の有効なサービスを提供しなくてはならないというジレンマがある。基本的には規範性を重視するとしても、どこまで幅広い記述性を加味すればよいのかという、規範性と記述性とのバランスの取り方、あるいは、ウェイトの

置き方が問題になってくる。

記述性を極端なまでに重視すると、「実際に使用されている言葉ならすべて変換可能にしよう」という立場をとることになるだろう。むろん、ユーザーの多様な変換要求に対して広範に対応できることは大切だが、間口を広くすればするほど、規範性の維持は困難になる。そのため、最近の仮名漢字変換システムでは、誤用が慣習化した表現を使おうとすると警告を表示したり、同音同訓語の使い分けを用例を用いて表示したりする補助的なガイド機能をもつものもある。その機能によって規範性の水準を維持しようとしているのである。

もちろん、このガイド機能を実現するためには、変換辞書の中に、規範性に関する判定情報を書き込んでおかなくてはならない。日本語の表記については、信頼に足る「規範」が明確でなく、多くの場合、「規範」と「慣用」と「俗用」との境界が曖昧なものになりやすい。そのため、この作業はきわめて厄介な作業である。変換辞書の制作者は、「規範性か記述性か」という難題に常に直面している。

六　「完成品」か「未完成品」か

通常、仮名漢字変換システムでは、ユーザーの運用結果を反映しながら、変換候補の並び順の入れ替えを行ったり、ユーザー単語の追加登録、共起情報の追加登録を行ったりしている。これらは順次、変換辞書に書き込まれていくことになるので、変換辞書の内容と編成は、使えば使うほど個別化し特殊化していく。また、付属の辞書ユーティリティーを用いれば、変換辞書の内容に対して、ユーザーが明示的に追加登録や変更や削除を行うこともできる。

そうしてみると、初期状態の変換辞書とは、いったい「完成品」なのだろうか。それとも「未完成品」なのだろうか。常に変化していくものであるならば、変換辞書はいつまでたっても「未完成品」であると言うこともできそうだ。しかし、これが制作者側の責任回避の言い逃れになっては困る。初期状態でどれほどの完成度を保証してくれているのか——これが、これからの仮名漢字変換システムの評価ポイントにならなくてはならないだろう。

特に、登録語彙の選択基準が恣意的であったり、読みや表記のバリエーションに整合性がなかったりするようなものに対しては、利用者として厳しい目を向けていく必要がある。一方、変換辞書の制作者には、日本語の言語生活と言語文化の一端を担っているという責任意識が必要であるし、何より

も、利用者に対する説明責任を忘れてほしくない。これまでの変換辞書にありがちだった「恣意性の隠蔽」は、利用者と制作者、どちらの立場からも安易に許容すべきではない。

ただし、誤変換がまさにそうであるように、仮名漢字変換システムの問題点の多くは、変換システムとユーザーの相互作用の中から発生する。利用者側の責任と制作者側の責任とは複雑に交錯する。実際の使用の場と、利用者との関わりの中にしか、仮名漢字変換システムの本質的価値（道具としての価値）はないわけだから、この点は致し方ない。

ここに胃袋があるということを意識させられるようでは、それは健康な状態ではない。胃袋はそれが「透明」な存在であるときが健康な状態なのだ。それと同様に、仮名漢字変換システムも「きちんと変換できて当たり前」でなくてはならない。その水準にいっそう近づいていくために、今後も変換辞書をめぐる「闘い」は続けられていくことだろう。

私たちはそろそろ、「変換辞書は文化財か消費財か」という、誠に興味深い問を意識し始めてもいいのかもしれない。

（たかもとじょうじ／日本語学）