語彙チェッカーを用いた読解テキストの分析

川村　よし子

１．はじめに

教師が教材を選択する場合、何を規準に教材を選んでいるのだろうか。おそらく、その選択の基準は、書かれている内容と文章の難易度によるという場合が多いであろう。では、その文章の難易度そのものはどのようにして判断しているのであろうか。

読解教材の難易度を決定する要素は漢字・語彙・文法・構文等いろいろあり、しかも各々の要素が複雑に絡み合っている。さらに、教材の内容自体も難易度に大きくかかわってくる。筆者は現在、それらの要素を個々に分解し、要素ごとの難易度をコンピュータで自動的に判定するためのシステムとして、「読解教材のレベル判定システム」の開発を進めている。本研究はそのシステムの一つとして開発した語彙のレベル判定システム「語彙チェッカー」を用いて行った読解テキストの分析結果の報告である。

語彙チェッカーは教材に含まれるすべての語彙の難易度を日本語能力試験の語彙リストを規準に判定するものである。このチェッカーを用いることにより、教師も学習者も自らが選んだ教材内にどんなレベルの語彙がどの程度あるかを事前に容易に知ることが可能になる。さらに、このチェッカーは、語彙の難易度と教材自体の難易度との関連を研究するための有用なツールとして活用可能である。

２．先行研究

コンピュータを日本語の読解教育に用いる試みはすでに 1980 年代から始められ、平沢ほか(1992)、加納（1992)等にその開発の報告がある。また、市販の読解指導用教材としては"Understanding Written Japanese" や"Nihongo Tutorial System"等が提供されている。(大坪,1992）これらの多くは読解教材およびそれを用いたドリルの提供に主眼がおかれ、いかに学習させるかという視点から開発され、いわゆるＣＡＩ(Computer-Assisssted Instruction)として作られたコンピュータ教材であり、教材の選択や提示の順番、さらにそれを用いた学習ドリル等がシステムにあらかじめ組み込まれている物が多かった。

これに対して、近年、ＣＡＬ(Computer-Assisted Learning)あるいはＣＡＬＬ(Computer-Assisted Language Learning)という呼び名で示されるように、学習者の視点からの教材開発が求められる（Higgins,1982 岡本,1983）とともに、学習者自身が教材を選択し、自らのニーズとレベルにあった教材を自由に選ぶことのできるシステムの構築が望まれるようになり、教材センター、教材バンクという考え方が提唱されるようになった（＊注１）。しかし、現実にこうしたシステムを維持するには大量のリソースをあらかじめ準備し、さらに常に新たなものを加えて提供し続ける必要がある。また個々のリソースを教材として有効に活用させるための学習支援環境も整える必要もある。こうした事情から、その必要

性は認識されつつも、ごく限られた教育機関において実現されているに過ぎなかった。

ところが、近年、インターネットの普及に伴い、教材として活用できる可能性のある大量な情報を電子文書の形で簡単に入手しうる状況になってきた。しかも、これらの中には、その時々の最新の情報が提供されている場合も増えてきた。また、教室活動等で日本語教師が提供する教材自体も準備段階でワープロ等で入力され、電子化情報として保存されていることも多い。つまり、今や大量の電子化リソースを教材として用いうる状況にあり、単語の意味や漢字の読みなどを教えてくれる学習支援システムさえ整えば、学習者が自由に自分の読みたい教材を選び、自ら学んでいくことの出来る環境ができあがりつつあるわけである（＊注２）。小森(1993)はコンピュータの検索機能を語法指導や作文指導に活用する方法を紹介している。また、寺・北村・落水によって開発された自動的に辞書引き作業をしてくれる辞書機能を備えた「読解支援システムＤＬ」（寺ほか,1996）や、熟語の利用頻度に応じてルビ振りを自動的に行う漢字学習支援システム「JUPITER」（越智,1997）等は、こうした時代の要請に応えたものといえよう。畑佐(1991)のいう本来の意味での「学習者のための道具」としてコンピュータが活用される時代に入りつつあると言えよう。

以上のような一連の流れをうけて、筆者は、電子化情報として提供されている個々の文章が、日本語学習者にとってどの程度の難易度かを自動的に判定する「読解教材のレベル判定システム」の開発を進めている。コンピュータによる語彙のレベル判定の試みは、英語教材に関してはなされている（Nation ほか,1996)ものの、日本語に関しては構文解析が難しいこともありこれまでは行われていなかった。学習者あるいは教師がこのシステムを用いることにより、電子化されたリソースを教材として有効に活用することが容易になる。

このレベル判定システムは、漢字・語彙・文法という３要素をそれぞれ個々にレベル判定するシステムで、３つのレベルチェッカー（漢字チェッカー・語彙チェッカー・文法チェッカー）から成り立っている。漢字チェッカーに関しては川村(in　press)で詳述している。また、川村(1998)では、語彙チェッカーの開発報告を行った。本論では、実際にこの語彙チェッカーを用いたテキストの分析に焦点を当て、語彙チェッカーの機能と、それを用いた教材分析の実際について述べることにする。

３．語彙チェッカーの仕組み

３．１語彙チェッカーの仕組み

語彙チェッカーは、教材に含まれるすべての語彙のレベルを日本語能力試験の語彙レベルを規準に判定するものである。語彙チェッカーはＣＧＩ化してホームページ上で動くようにしてあるため、インターネットに接続できる環境にあれば、自由に利用できる。さらに、電子化されたテキストファイルであれば、ワープロ等で入力された教材でも、ＯＣＲで電子情報化された教材でも、インターネット等で提供されている電子情報でも、どんな文書でもすべて受けつける。そして与えられたテキストに対して、コンピュータが次の作業を自動的に行い、テキストに含まれるすべての語彙のレベルを判定する。

①与えられたテキストの形態素解析を行う

②各形態素をレベル別語彙リストに照合する

③テキスト内の語彙にレベル表示を行う
④テキスト内語彙のレベル別分類表を作成する
⑤語彙のレベル別含有率を算出する
以下に語彙チェッカーの行う作業行程を詳しく述べることにする。

３．２テキストの形態素解析

入力されたテキストを形態素解析し、語句ごとに区切りをいれる作業には、奈良先端科学技術大学院大学の松本研究室によって開発された形態素解析ツール「茶筌」を用いた（＊注３）。英語のように、単語ごとにスペースで区切られていれば辞書との照合は比較的容易だが、区切りが示されることの少ない日本語では、コンピュータに文中の個々の単語を認識させるためには、まず、あらかじめ単語ごとに区切る作業をしなければならない。形態素解析ツールである「茶筌」を用いることにより、入力された文は形態素ごとに切り分けられ、読み、品詞情報等が自動的に識別される。

例えば本文の第一文を茶筌にかけてみると、次のように分析される。

【第一文】

教師が教材を選択する場合、何を規準に教材を選んでいるのだろうか。

【茶筌による分析】

教師,教師,きょうし,名詞,普通名詞,*,*,1000
が,が,が,助詞,格助詞,*,*,100
教材,教材,きょうざい,名詞,普通名詞,*,*,1000
を,を,を,助詞,格助詞,*,*,100
選択,選択,せんたく,名詞,サ変名詞,*,*,1000
する,する,する,動詞,*,サ変動詞,基本形,700
場合,場合,ばあい,名詞,副詞的名詞,*,*,990
、,、,、,特殊,読点,*,*,1000
何,何,なに,名詞,普通名詞,*,*,1000
を,を,を,助詞,格助詞,*,*,100
規準,規準,きじゅん,名詞,普通名詞,*,*,1000
に,に,に,助詞,格助詞,*,*,100
教材,教材,きょうざい,名詞,普通名詞,*,*,1000
を,を,を,助詞,格助詞,*,*,100
選んで,選ぶ,えらんで,動詞,*,子音動詞バ行,タ系連用テ形,1000
いる,いる,いる,接尾辞,動詞性接尾辞,母音動詞,基本形,100
のだろう,のだ,のだろう,助動詞,*,ナ形容詞,ダ列基本推量形,100
か,か,か,助詞,終助詞,*,*,100
。,。,。,特殊,句点,*,*,1000
EOS,,,,,,,,,,,,

元の文が単語ごとに区切られ、基本形、読みが表示されるのみならず、品詞分類、さらに、活用形までが自動的に分析されるわけである。各行の最後の数字は、単語の区切りを決定するための重み付けである。

ただし、この例でも明らかなように、茶筌の分類は従来の国文法や日本語文法の品詞分類とは多少異なった部分が存在する。例えば、「選択する」は「サ変動詞」ではなく、「サ変名詞」の「選択」と「動詞」の「する」との２語として解析されている。また、補助動詞的に用いられた「いる」は「接尾辞」として扱われている。そこで、形態素解析ツールとして茶筌を用いる場合には、こうした特殊な分析に留意し、これに対応する形で開発を進める必要がある。

語彙チェッカーではこの茶筌の解析で得られた語彙の一つ一つをレベル別語彙リストと照合することになる。

３．３レベル別語彙リストとの照合

照合先のレベル別語彙リストは、日本語能力試験の１級から４級までの各級の語彙表をもとにして作成した。情報が欠けている場合、例えば、ある語彙が平仮名表記で書かれていた等の場合には、その語彙の特定が難しい。そこで級の判定に複数の候補が考えられる場合は、下の級を優先することにした。

日本語能力試験の各級の語彙表をもとにしたレベル別語彙リストの作成は次のようにして行った。

①読みの入力

各単語の読みをひらがなで入力する。この場合、カタカナ語のように、音の表記に複数の表記が考えられる場合は双方を入力した。（例：ばいおりん／ｳﾞぁいおりん）

②表記の入力

各単語の漢字仮名交じり表記を入力する。取り扱いに問題のある送りがなや、補助的に用いられる動詞等、複数の表記が存在する場合、考えられるすべての表記を入力した。（例：よりかかる／寄掛かる／寄りかかる／寄り掛かる／寄掛る／寄り掛る）

③形容動詞とサ変動詞の扱い

茶筌の辞書と同一の形式にするため、形容動詞は「だ」のついた形、また、サ変動詞は「する」をとった形で入力した。日本語能力試験の語彙表には品詞名の記載はなく、また、形容動詞に関しては、ごく一部の単語を除き、語幹が示されているのみである。そのため、語彙表にあげられている単語が名詞か形容動詞かは明らかではない。そこで語彙リストの作成においては、各々の語を名詞として登録した上で、形容動詞として用いられる可能性のあるものに関してはすべて形容動詞としても登録した。また、副詞として用いられる可能性のあるものに関しては、副詞としても登録した。

④品詞名の入力

各単語の品詞名を茶筌の品詞分類に基づいて入力する。上述のように従来の品詞分類と異なっている場合も多い。特に、従来の国文法でいわゆる助動詞に属するものの分類は複雑である。終止形に接続するものは「助動詞」のままだが、その他の活用形に接続するものは「接尾辞」として分類されている。また「だ／です／である」は「判定詞」として分類されている。

⑤文法項目として取りあげられている語彙

日本語能力試験において、文法項目として取りあげられている語彙に関しては、語彙表に取りあげられていないものも多い。（例えば、助詞や助動詞などはその大半が語彙表にはのっていない。）語彙チェッカーの級別語彙リストにはこれらの語彙も級ごとに付け加えることにした。

以上の作業の結果、１級から４級までのレベル別語彙リストが完成した。

語彙チェッカーは、この語彙リストに、上述の形態素解析の結果得られた語彙を照合することによって、入力されたテキストのすべての語彙に対してレベル判定を行う。この場合、例えば、テキストがひらがなで表記されている等の理由で語彙の特定が難しく、合致する語彙が複数候補考えられる場合には、下の級を優先する。さらに、４級から１級までのすべてのリストに合致しない語彙は、すべて「表外語彙」として扱うことにした。

３．４　ＣＧＩによる開発

語彙チェッカーにおいて形態素解析を行っている「茶筌」は、ＵＮＩＸ上で動作するツールである。従って、可能な限り多くのユーザーに利用可能な環境を提供するためには、語彙チェッカーのシステム全体をインターネットを介して利用できるようにする必要がある。そこで、本研究ではＣＧＩを用いてネット環境の中で利用できる仕組みとして構築することにした。その結果、利用者は、インターネットを使える環境にさえあれば、自由に語彙チェッカーを利用することが可能になった。（＊注４）

利用者は、語彙のレベル判定を行いたいテキストを、ホームページ上の入力画面（＊図１）に入力し、「Ｓｅｎｄ」ボタンをクリックして、データーを送り出しさえすればよい。他のホームページ等からの電子情報の切り張りも可能である。その後の解析・判定作業はすべてコンピュータが自動的に行ってくれる。例えば、資料１の文をテキストとして語彙チェッカーにかけてみよう。＊図２はその解析結果の表示画面である。画面に入りきらない情報は、スクロールバーを使ってスクロールして見ることができる。

次章で、語彙チェッカーを用いることによって実際にどのような分析を行うことができるのか見ていくことにする。

４．語彙チェッカーによる語彙のレベル判定

４．１　語彙チェッカーによるテキストの分析

語彙チェッカーはまず、テキストの形態素解析を行い、語彙ごとに区切りをいれる。図２の左上のテキストも、その区切りに応じてスペースが挿入されている。これを見ることによって、どのような解析が行われたのかが明らかになる。上述したように、茶筌の解析では単に区切りをいれるのみならず、読みや品詞情報等も分析されている。その分析結果は画面には見えないが、テンポラリーファイルに保存され、級別リストとの照合にも用いられる。

テキストの形態素解析の結果抽出された語彙を級別リストと照合した結果は、左上の画面に級ごとに色分けして表示される。ここでは印刷の都合上、１級以上の語彙（１級の語彙および級別リストに含まれなかった表外語彙）を太字で表示してある。実際には級ごと

に異なった色で区別して表示できるようになっている。また、学習者のレベルにあわせて（例えば、この例のように、１級以上と２級以下とを区別して）色分けすることも可能である。

４．２　テキスト内語彙のレベル別分類表の作成

図２の右上の表は、テキスト内の語彙を級別に分類したものである。テキストに含まれるすべての語彙について級別語彙リストと照合した結果が表示されている。さらに各語彙の使用頻度も示されている（各語彙の右の数字がそれである）ので、どのレベルのどのような語彙がどのくらい用いられているかが即座にわかる。

例えば、資料１のテキストに含まれる１級の語彙は次の通りである。

背景　はかない　象徴　冒頭　題する　なにより　告げる　現行　版　すえる　かかげる　なめらかだ　移行　修飾　描写　剥ぐ　結びつく　独創　結合　むしろ　闇　視点　名づける　～がたい　保つ　すえる　一変　ふさわしい　きっぱりと　添える　空間　耐える　映像　にもかかわらず　読者　いとなむ

また、１級から４級までのすべてのリストに含まれなかったものは「表外語彙」として分類されている。資料１のテキストに含まれる表外語彙は次の通りである。（＊注５）

雪国　川端　康成　印象的だ　内の人　一刻　二重　夕闇　おぼろだ　融ける　この世　ただなか　野山　ともす　こういう　ときには　触感　なまなましい　島村　書き出す　一文　切り離す　景　不要だ　通り抜ける　視線　連体　切り落とす　感性　際立つ　芥川　末尾　下人　またたく　自体　奇　てらう　異様だ　はっと　なかに　言える　ぼうっと　緊迫　深さ　厚み　奥行き　そういう　知覚　融合　読める　改稿　際　すっくと　官能　回想　車窓　映ずる　修飾語　暗がり　白一色　ひろがり　華麗だ　無為　徒労　駒子　葉子　彼岸　あの世　なぞらえる　書き出す　ありげだ

教師が新たな教材を用いる場合、テキスト内の重要語や新出語を語彙リストとして示すことも多い。その際に、このレベル別分類表を活用すれば、学習者に提示する必要のある語彙が明らかになる。また、学習者自身が語彙チェッカーを用いれば、自分にとっての未習語彙のリストを簡単に作成することができる。さらに、このレベル判定を辞書システム、あるいは、ふりがなシステム等と併用すれば（＊注６）、学びたい内容のテキストを自らのレベルにあわせて即座に教材化することも可能になる。

４．３　テキスト内語彙のレベル別含有率の算定

図２の下の表は、テキスト内の語彙を級ごとにカウントし、語彙の総数に対する級ごとの含有率を算出したものである。語彙の難易度と教材の難易度とはそのまま一致しない場合もあるが、学習者にとって未習の語彙が多ければ、教材がそれだけ難しくなることは明らかであろう。

資料１のテキストは実際に 1997 年度の研修日本語の読解のクラスで教材として用いた

ことがあるが、日本語能力試験の１級レベルの学習の終了した学生にとって、この教材には未習の語彙が 139 語（14.8%）含まれていることになる。また、２級終了レベルの学生にとっては、未習の語彙は 186 語（19.9%）となる（＊注７）。

この語彙チェッカーの開発の目的の一つは、テキストの難易度と語彙の難易度との関連を調べることにある。テキストの難易度と級ごとの語彙の含有率にはどのような関係があるのか。現実のテキストには、どのようなレベルの語彙がどのくらい使われているのか。未習の語彙が何%位であれば、教材として利用可能なのか等々、教材の難易度と語彙の難易度との関連を調べるための基礎的データの分析には有用なツールとなるはずである。

５．語彙チェッカーの分析の問題点

語彙チェッカーは、上述したように、形態素解析を「茶筌」に負っている。語彙チェッカーによる語彙のレベル判定が正しく行われるためには､茶筌による解析が正確に行われ、さらにその解析結果と級別語彙リストとの照合がうまく行われる必要がある。

第３章（3.3）で述べたように、語彙リストの作成においては、茶筌の解析結果を最大限生かしうるように配慮はしたものの、語彙チェッカーは、茶筌の解析能力（＊注８）以上の解析能力は持たない。また、形態素解析のための辞書の登録方法は一般の辞書とは異なっている部分もある。こうしたことから、語彙チェッカーにはレベル判定上の問題がいくつか生じる。語彙チェッカーによる分析に、どのような問題が生じ、それに対してどのような対処をすればよいのかを以下に論じることにする。

５．１　語彙の区切りの誤り

語彙チェッカーがどの程度正確に解析を行えるかを資料１（総文字数 1899 字）を元に調査した。語彙チェッカーは資料１を 937 語の語彙に区切っている。おのおのの語彙の区切りが正しいかどうか、詳しく調べたところ、語彙の区切りの誤りは 25 箇所であった。ところが、このうち、13 ヵ所は旧仮名による表記が原因であり、5 ヵ所は常用漢字以外の漢字が用いられていることが原因であった。そのため、実質的な区切りの誤りは７ヵ所ということになる。（＊注９）

５．２　漢字を含んだ語彙の解析

漢字は延べ 555 字用いられていた。これらの漢字を含んだ語彙の解析結果を調べたところ、区切りの誤りが５カ所見つかった。その誤りは、いずれも、漢字が２字以上連続している所でおきていた。（＊注１０）各々の誤りの発生の原因とその解決方法は次の通りである。

・「車内の人」→「車・内の人」

茶筌の辞書に「内の人」はあったが､「車内」という言葉が登録されていなかったために起こった誤りである。この問題は、「車内」「室内」「館内」等、接尾辞を含む語を辞書に登録することによって、解決できる。ただし、こうした登録を行う際には、どのような単語をどういう基準で登録するか一定のルールを決める必要がある｡また、

「～内」を含む語を登録するのであれば、「～外」を含む語も登録する必要が生じる。

・「導入部」→「導・入部」

茶筌には「導入」も「入部」もサ変動詞（茶筌では「サ変名詞」と呼ばれている）として登録され、さらに、「導（しるべ／名詞）」「部（ぶ／名詞）」も登録されている。そこで、「導・入部」と区切る可能性と「導入・部」と区切ることも可能性は同程度にある。こうした区切りを正確に行わせるためには、茶筌に「部」を接尾辞として登録する必要がある。

・「視点人物島村」→「視点・人・物・島村」

本来ならば「視点・人物・島村」と区切るべきものである。茶筌の辞書に「人物」という単語も登録されている。ところがこの例の場合、「人」も「物」も「接尾辞」としても登録されている単語である。そのため、「人物」の前に「視点」という名詞がおかれているので、各々接尾辞として解析した結果生じた誤りである。こうした誤りを避けるには「視点人物」という単語を登録するしかない。

・「芥川龍之介」→「芥川・龍・之・介」／「羅生門」→「羅・生・門」

固有名詞に関しては、辞書に登録するのが唯一の解決策である。ただし、登録語彙が増えれば増えるほど、解析に時間がかかることになる。また、固有名詞をすべて登録することは事実上不可能である。語彙チェッカーによってどのような内容の文章を分析しようとするのかによって、登録すべき語彙も異なってくる。また、固有名詞に関してはこうした解析上の問題が生じてくることさえ承知していれば、解析結果自体に大きな問題はないとも言える。これらのことを考慮した上で、必要であれば茶筌の固有名詞辞書に登録することが可能である（*注１１）。

以上が漢字を含む語彙の誤りの分析とその解決策である。上記の他にも漢字語彙の解析上の問題は生じうるが、基本的には辞書の整備によって解決するものが多い。

５．３　ひらがな表記の多い文の解析

語彙チェッカーは、ひらがな表記が長く続く場合にも、正確な解析が行えないことがある。資料１の解析では、「ただ人をはっとさせるだけではない」「むしろきちんととらえられている」等の文は、ひらがなが多用されてはいるが、「ただ・人・を・はっと・させる・だけ・で・は・ない」「むしろ・きちんと・とらえ・られて・いる」のように正確に解析される。ところが、図２にも示されているように、テキスト中の「登場人物と背景とがなんのかかわりももたない」という文は「登場・人物・と・背景・とが・なんの・か・かわり・も・もたない」と分析されている。「かかわり」がひらがなで表記されていたため茶筌が「かかわる」をうまく抽出できなかったのである。また、資料１の最終部分の「まぶしさをともなって読者をひきつける」も「まぶし・さ・を・と・も・なって・読者・を・ひきつける」のように「ともなう」を抽出することは難しい。いずれの解析ミスも、「か」「と」「も」を助詞として区切って分析してしまったために生じたミスである。

茶筌は、ひらがな部分に助詞の可能性のある文字が存在すると、そこで、ひとまず、区切りをつくってしまう傾向がある。そして、残りの部分をうまく辞書に当てはまる形で区切ることができれば、解析を完了してしまう。こうした点はコンピュータによる自動解析

においては、目下のところ避けがたい分析ミスだと言えよう。

そのほかに、資料１で解析を誤った箇所は「白い雪のひろがりがぼんやりと見える」である。これは「白い・雪・の・ひろがり・が・ぼん・やり・と・見える」と解析してしまっている。この解析の誤りは茶筌の辞書に副詞「ぼんやりと」が登録されていなかったことによるものであり、辞書にこの語を付け加えることによって解決可能である。

５．４　旧仮名・旧字体の扱い

旧仮名（歴史的仮名遣い）および旧字体（常用漢字表以外の漢字）が用いられた場合、語彙チェッカーでは解析できないことが多い。

例えば、資料１は『雪国』からの引用で始まっていた。表記は原文を尊重して歴史的仮名遣いで書かれている。そのため、語彙チェッカーは、「あつた」「なつた」「止まつた」は、分析できない。また、文中の『夕景色の鏡』からの引用部分にある「融け合ひながら」も「融け・合・ひなが・ら」と分割し、「ひなが」という全く別の単語を抽出してしまう。

常用漢字に含まれていない漢字についても同様の問題が発生する。文中に用いられている「妖」「顫」は常用漢字以外の漢字である。そのため、これを含んだ「妖しい」「顫える」は分析できず、「妖・し・い」「顫・える」と区切ってしまうため、本来は１語として扱うべきものが２語あるいは３語として数えられてしまうことになる。

語彙チェッカーを用いる場合には、これらのことに留意する必要がある。まず、歴史的仮名遣いに関しては現在のところ語彙チェッカーでは解析できないので、テキスト自体を現代仮名遣いに直してから、分析する必要がある。これに対して、常用漢字以外の漢字に関しては、あらかじめ、茶筌の辞書にそれを含んだ単語を登録しておけば、解析できるようになる。例えば、「闇」や「梯」は常用漢字ではないが、茶筌の辞書には「闇（やみ）」「梯子（はしご）」という表記で、名詞として登録されている。そのため、これらは問題なく正確に解析される。語彙チェッカーを用いて様々な電子化文書を解析しようという場合には、こうした一般に用いられることの多い表記は、前もって、茶筌の辞書に登録しておく必要があろう。

５．５　複合語の扱い

茶筌の辞書にはすでに１５万語以上の語彙が登録されている。ところが、与えられた文から単語の切り出しを正確に効率よく行うために、茶筌の辞書には普通の国語辞典とは異なる基準で語彙が登録されている。一般に、文章の解析において、辞書に登録されている単語が長ければ長いほど解析の間違いは少なくなる。そのため、茶筌の辞書に見出し語として登録される語彙の中には、いわゆる意味の最小単位としての単語ではなく、単語がくみ合わさってできあがっている複合語も含まれているのである。

いっぽう、語彙チェッカーの語彙リストには、複合語で登録されているものは少ない。（注１２）そのため、構成要素の各単語に分割したときは級別語彙リストにのっている語であっても、複合語として１語で切り出されてしまうと、語彙チェッカーでは表外語彙として分類されることになる。例えば資料１のテキストでは次のような複合語が表外語彙になっている。

動詞＋助動詞　：言える／読める

動詞＋動詞　：見つめる／書き出す／切り離す／通り抜ける／ひきつける／言い切る
名詞＋名詞＋助詞　：白一色の
連体詞＋名詞　：この世／あの世

まず、動詞の可能形に関しては、茶筌では別の動詞として登録されているため、この問題が生じている。語彙チェッカーの語彙リストに新たに一連の可能形を登録する必要がある。

可能形以外の複合語に関しても、それを構成している単語の各々は１級から４級の語彙リスト内の単語であり、級別語彙リストに複合語が載っていないからといって、自動的に表外語彙と判定されてしまうのは問題である。ところが、これらの語の中には、意味がもとの単語の意味とは微妙にずれているものも多い。例えば、上記の中でも、「この世」と「あの世」は、新出語彙として扱っていいものである。また、動詞が補助動詞として用いられた場合にも、派生的な意味が生じることがある。語彙チェッカーが表外語彙として分類した語彙の中に、既習語で構成された複合語が含まれうるということさえあらかじめ承知していれば、むしろ、意味の理解の際に留意できるというメリットはあるとも言える。

こうした複合語が表外語彙と判定されてしまうのを避ける方法は２つある。一つは、語彙チェッカーの語彙リストにこれらの複合語を登録する方法であり、もう一つは茶筌の辞書から複合語を削除する方法である。いずれの方法をとるべきかに関しては解析能力との関係をも含めて検討していく必要がある。

５．６　接尾辞・接頭辞の扱い

接尾辞・接頭辞に関しても複合語と類似の問題が存在する。茶筌の辞書に、接尾辞・接頭辞のついた単語が登録されていれば、茶筌はこれを１語として切り出す。ただし、ここで、大きな問題となるのは同じ接尾辞あるいは接頭辞がついていても茶筌の辞書に登録されている場合と、されていない場合とがあるということである。

例えば、例２のテキストには「雪の深さと闇の厚み」という表現があるが、その中の「深さ」「厚み」はともに１語として切り出されて、表外語彙とし分類されている。ところが、同じテキストの中にある「白さ」「まぶしさ」「はかなさ」等は、おのおの形容詞の「白い」「まぶしい」「はかない」と接尾辞「さ」とが別々に２語として切り出され、不統一が生じている。

接頭辞に関しても同じことがいえる。例えばテキスト中の「非連続感」は図２が示すように「非・連続・感」のように３語として扱われているが、「非合法」「非合理」「非公式」等は１語として扱われる。また、「不親切だ」「不心得だ」「不行き届きだ」は１語として登録されているのに、「不確定だ」「不統一だ」「不似合いだ」等は登録されていない。

接尾辞・接頭辞のついた単語をすべて登録することは不可能であり、解析結果に統一性を持たせるには、逆に、接尾辞・接頭辞のついた単語を茶筌の辞書から削除するしかない。削除にあたっては、それによって茶筌の解析能力が落ちることはないかどうかを十分に検討した上で、対処しなければならない。また、級別語彙リストには基本的に接尾辞・接頭辞のついた単語は登録されていないので、両者の間の整合性も含めて考えている必要があろう。

以上、語彙チェッカーの分析の問題点を論じてきた。現在のところ、茶筌の解析能力は98.5 %だというデータがある（注８参照)。また、本研究の分析結果では、資料１の引用箇所を除いた部分の解析の適合率は 98.0%であった。（＊注１３）つまり、通常の文を入力すると、200 語につき３箇所ないし４箇所、解析の誤りが生じてしまうことになる。しかし、上述のように、辞書の整備（茶筌の辞書に必要な語彙情報を付け加える、語彙登録の基準を統一する、接尾辞・接頭辞の扱いを再検討する等）をさらにすすめることによって、茶筌および語彙チェッカーの解析能力を高めていくことが可能なはずである。

６．今後の課題

学習者にとっても、教師にとっても、テキストそのものの難易度の判定を自動化できれば教材の選択が容易になるばかりでなく、学習者のレベルにあった適切な教材の選択が可能になる。難易度判定は、ある程度の相対評価は出来ても絶対評価は難しいものがある。そこで、教材としての難易度とテキストに含まれる語彙の級別の比率とをどう関連づけることができるかに関して、現在、この語彙チェッカーを用いて調査を進めている。その調査結果は、語彙チェッカーを含むレベル判定システムに反映させ、テキスト自体の難易度の判定の自動化をはかる予定である。

また、語彙チェッカーの場合、文要素の解析自体は「茶筌」に負っている。茶筌の解析の適合率が高くなれば、語彙チェッカーの分析能力もそれだけ高まる。語彙チェッカーの活用で明らかになる解析上の問題点をフィードバックすることで、茶筌自体の解析精度の向上に貢献できればと考えている。

インターネットの普及に伴い、最新の情報が電子化文書という形で手にはいるようになった今、日本語学習のためのリソースとしてもこれらを有効に活用する必要がある。そのためにも語彙チェッカーを含むレベル判定システムを、学習支援のためのツールとして、日本語教師及び日本語学習者が利用しやすい形で提供していく必要がある。本研究の成果をホームページ上で公開し、ツールとして活用してもらうとともに、実際の活用で生じた問題点をフィードバックしてもらい、解析能力と使いやすさを高めていきたいと考えている。

謝辞：本研究をまとめるにあたって、奈良先端科学技術大学院大学の松本裕治先生、静岡大学の北村達也氏に貴重なご助言を得た。また、語彙チェッカーの開発には東京大学理学部の保原麗氏の協力を得た。ここに記して、感謝の意を表したい。

参考文献

浅木森利昭(1994)『マルチメディアを利用した日本語教育支援システムの開発　平成５年度科学　研究費補助金研究成果報告書』

大坪一夫(1992)「日本語教育でのコンピュータ利用の過去、現在と未来」『日本語教育』78, pp.9-19.

岡本敏雄(1983)［ＣＡＩからＣＡＬへ」『ＣＡＩ学会誌』Vol.3, No.2-3, pp.21-23.

越智洋司(1997)「電子化された日本語文書を教材とした漢字学習システム」JCET(http://www-yano.is.tokushima-u.ac.jp)

加納千恵子(1992)「ＣＡＩを利用した授業研究の可能性－－日本語読解支援システムの開発と授業分析」『日本語教育』78 号, pp.131-140.

川村よし子(1998)「 読解のためのレベル判定システムの構築 －語彙チェッカーの開発と活用－」『日本語教育方法研究会誌』Vol.5, No.2.

川村よし子(in　press)「漢字の難易度判定システム『漢字チェッカー』を用いたテキストの分析」『東京国際大学論叢』第 59 号.

小森早江子(1993)「中級日本語教育のための検索プログラムの利用法」『コンピュータ利用の外国語教育』英潮社, pp.187-194.

芝野・川村・田崎(1998)「コンピュータ支援によるオーディオ及びビデオを用いた一般教育法及び語学教育法の研究」『東京国際大学論叢』第 57 号, pp.97-111.

鈴木庸子(1998)「日本語学習者を対象とした読書支援システムの開発」『文部省科学研究費補助金 1997 年度研究成果報告書』

寺・北村・落水(1996)「WWWブラウザを利用した日本語読解支援システム」『日本語教育方法研究会誌』Vol.3, No.1, pp.10-11.

東京外国語大学留学生日本語教育センター編(1998)『中・上級社会科学系読解教材テキストバンク』東京外国語大学留学生日本語教育センター.

日本語能力試験企画小委員会編(1993)『日本語能力試験出題基準（外部公開用』国際交流基金・日本国際教育協会.

畑佐一味(1991)「日本語教育におけるコンピュータ利用－米国からの一考察－」『日本語教育』74 号, pp.162-171.

平澤・渋井編(1992)『日本語ＣＡＩの研究』桜楓社

松本・北内・山下・今・今村(1997)日本語形態素解析システム「茶筌」ver 1.0 使用説明書,NAIST TechnicalReport,NAIST-IS-TR97007.

Higgins,John(1982)"Computer Assisted Language Learning,"Language Teaching, Vol.16, No.2, pp.102-114.

Nation,P. et al.(1996)Using texts to sequence the introduction of new vocabulary in an EAP Course, RELC Journal27,2:1-11.

（資料１）

国境の長いトンネルを抜けると雪国であつた。夜の底が白くなつた。信号所に
汽車が止まつた。　　――川端康成『雪国』

この作品の初めの部分、しだいに暮れてゆく汽車の窓に顔が映る、その娘の顔の奥を夕景色が流れる、あの場面は印象的だ。外が暗くなると窓ガラスは鏡に変わる。その少し前、透明なガラスを通して車外の景色を見せながら、同時に、鏡の働きをして車内の人を映しだす一刻がある。この作家はそれを、登場人物と背景とがなんのかかわりももたない映画の二重写しと見る。じっと見つめていると、「人物は透明のはかなさで、風景は夕闇のおぼろな流れで、その二つが融け合ひながらこの世ならぬ象徴の世界を描いて」いるように見える。そして、「娘の顔のただなかに野山のともし火がとも」り、それが瞳に重なった瞬間には、その妖しい美しさに胸が顫えるのである。

『雪国』の冒頭を飾るこの場面は、こういう感動をのせて、まず『夕景色の鏡』と題する短編として発表された。そのときには「濡れた髪を指でさはつた。――その触感をなによりも覚えてゐる、その一つだけがなまなましく思い出されると、島村は女に告げたくて、汽車に乗つた旅であつた」と書き出され、それにつづくこの場面も少し違って、現行版の初めの二行に相当する部分は、「国境のトンネルを抜けると、窓の外の夜の底が白くなつた」となっていたらしい。

逆にいえば、『夕景色の鏡』の一文を『雪国』では二文に切り離し、それを冒頭にすえたことになる。窓外の景が〈雪国〉として冒頭にかかげられた。そのために不要になった「窓の外の」という説明を削除し、第二文ではすでに窓ガラスを通り抜けたその視線の先を追う。ここで「窓の外の」となめらかに移行する連体修飾が切り落とされた結果、「夜の底が白くなつた」という感性のまぶしい描写が一文として独立性を増し、いっそう際立つこととなる。

芥川龍之介の『羅生門』の末尾にも「下人は、剥ぎとつた檜皮色の着物をわきにかかへて、またたく間に急な梯子を夜の底へかけ下りた」という例があるから、「夜の底」という結びつき自体は川端康成の独創とはいえない。しかし、その語結合がさらに「白くなる」と結びついた全体は、この作家の前におそらく誰も書かなかっただろう。

奇をてらったようにも見えるこの異様な表現は、ただ人をはっとさせるだけではない。「暗いなかに、降り積もった白い雪のひろがりがぼんやりと見える」とでも言えそうな現実が、むしろきちんととらえられている。しかも、「闇の奥に白さがぼうっと感じられた」というふうに薄く流れてしまわない緊迫感がある。雪の深さと闇の厚みという現実の奥行を写しだすことができたと言っていい。と同時に、そういう現実を知覚におさめた感動――驚きと期待といくばくかの不安が融合した、視点人物島村の何ものとも名づけがたいそういう感情が、ことばの奥を流れているようにも読める。

改稿の際に「夜の底が白くなつた」という部分が切り離され、すっくと立つ。そのとき、新しい冒頭文もそれによって独立性を増し、適度の非連続感を保ちながら、あざやかに立ち上がった。『夕景色の鏡』の官能的に誘いこむ回想の導入部を切り落とし、『雪国』の新しい冒頭文は、まさに〈雪国〉の世界を正面にすえて、象徴的に幕を開く。

トンネルを一つ越えるだけで車窓に映ずる風景が一変する。新鮮な驚きは、やはり、それだけがすっきりと立つ一文で描かれるのがふさわしい。「国境の長いトンネルを抜けると雪国であつた」――きっぱりとそう言い切ったのは、そのためではなかったろうか。

雪国というとらえかたにも、それをこのようにいきなり投ずる表現の姿にも、視点人物島村をとおして作者の感動が映っているような気もする。改稿時にトンネルに新しく添えた「長い」という連体修飾語も、空間的な距離の強調であるよりは、闇に耐えている心理的な時間の強調であったかもしれない。「長いトンネル」という暗がりを通り抜けた瞬間、思いもかけず、〈雪国〉という白一色の世界が立ち現れる。闇の底にぼうっと沈む白いひろがり――それはけっして華麗な映像ではない。にもかかわらず、このイメージの展開は一瞬のまぶしさをともなって読者をひきつける。

無為徒労の現実の生活がいとなまれるこちらの側の世界と、駒子や葉子の住む向こう側の世界――長いトンネルの手前と先を、此岸と彼岸、つまり、この世とあの世になぞらえる深読みがある。そう読んでもおかしくないほど、この書き出しの表現がなにやら意味ありげな姿で立っていることは疑えない。

（中村明『文章をみがく』による）

注

１　国立教育研究所の「CASTEL-J」（浅木森 1994）やこれをもとに開発された「新書ライブラリー」（鈴木 1998）さらに東京外国語大学による「社会学系読解教材テキストバンク」等はこの理念のもとに開発されたものである。

２　芝野・川村・田崎(1998)「コンピュータ支援によるオーディオ及びビデオを用いた一般教育法及び語学教育法の研究」『東京国際大学論叢』第 57 号, pp.97-111.

３　これは http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html のホームページ上でフリーウェアとして公開され、自由にダウンロードして用いることが許されているソフトである。こうした貴重なツールが公開されていることの意義は大きい。

４　「語彙チェッカー」は現在、http://www.tiu.ac.jp/~kawamura で、学習者および教師が自由に活用できる形で、「漢字チェッカー」とともに公開している。

５　これは、語彙チェッカーによる解析の結果抽出された表外語彙のリストだが、複合語等で一語として切り出されているために表外に分類されているものが含まれている。また、解析に問題のあるものは、このリストには載せていない。いずれも第５章で詳しく扱う。

６　例えば、寺・北村・落水(1996)の「読解支援システムＤＬ」と組み合わせて用いることにより、個々の学習者にあわせた学習支援ツールの提供が可能となる。

７　この数字はあくまでも語彙チェッカーによる解析の結果の数字である。後述するように、旧字・旧仮名が用いられた場合、解析精度がおちるし、解析の誤りもあるので、資料１に含まれる表外語彙の実数はこれとは異なっている。

８　茶筌の解析の適合率は現在の所 98.5%とのことである。（日本語形態素解析システム JTAG ホームページ http://lambda.cipl.cae.ntt.co.jp による）

９　ここでいう区切りの誤りとは、その区切りによって生じた各要素が意味的におかしい区切り方をしている場合をいう。したがって、接尾辞・接頭辞等の区切りはあってもなくてもよいことにした。

１０　送り仮名を伴う語を途中で区切る等の誤りは生じていなかった。

１１　茶筌自体の辞書にも、すでに約３万５千語の固有名詞が登録されている。

１２　日本語能力試験の語彙表には、複合語であっても、登録されているものがある。そこで、今回の級別語彙リストの作成にあたっては、複合語や、接尾語・接頭語が付加された語等に関しては、語彙表にあるもののみリストに登録したが、日本語能力試験の語彙表自体も統一性のあるものに整え直す必要があると言えよう。

１３　JTAG の調査結果に比べて適合率が低いのは、本文中に表外字が４文字含まれていることにより解析の誤りが増えたものと考えられる。