語彙チェッカーを用いた日本語教科書の分析
Analysis of Japanese textbooks using the "Vocabulary Level Checker"

川村よし子（東京国際大学）

KawamuraYoshiko（Tokyo InternationalUniversity）

概要：The Vocabulary Level Checker (VLC) automatically compares all the words in the text with the wordsinthelistsfromthefourlevelsintheJapaneseLanguageProficiencyTestandshowsthelevelof eachword. It also makesa list of theactual words found at each level andgivesthenumberofoccurrences, showing thepercentage for each level. Inthispaper,eightJapanese textbookswere analyzedusing the VLC in order to examine the relationbetweenthesupposed levelofdifficultyoftextsandthelevel of vocabulary. This analysis reveals that the intermediate textbooks contain many unknown words for intermediate level students:30%ofthewordswouldbeunknowntointermediatelevel students, and 10% for advanced students.

１．はじめに

近年、インターネットの普及に伴い、日本語で書かれた電子情報が容易に入手可能になった。しかも、これらの情報の中には、最新のニュースも含まれ、日本語学習者の教材リソースとして十分活用できるものも多い。今や、学習支援ツールさえあれば、世界中の学習者が自ら教材を選び、学べる状況になりつつある。また、学習支援のツールも徐々にではあるが提供され始めている。寺・北村ら(1996)は、コンピュータが自動的に辞書引き作業をしてくれる「読解支援システム DL」を開発した。越智ら(1997)の開発した漢字学習システム「JUPITER」も、ルビ振り機能をもち、漢字のテスト問題を自動的に作成する。さらに、DL はインターネットを通して、学習者が自由に利用できる形で公開されている。また、学習履歴管理の機能（北村・川村ら 1998）も整備された。まさに、読解のための学習支援環境が整い始めているといえよう。

こうした一連の流れをうけて、筆者は、学習および教育のための支援システムとして、テキストの難易度を自動判定する「読解教材のレベル判定システム」（川村　1998）の開発を進めている。読解教材の難易度を決定する要素は多様であり、語彙・漢字・文法・構文等、各々の要素が複雑に絡み合っている。「レベル判定システム」は、こ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しようというものである。このシステムは語彙・漢字・文法の３つのレベルチェッカーによって構成され、個々のチェッカーで解析した結果を統合して、読解教材の難易度を自動的に判定することを目指している。

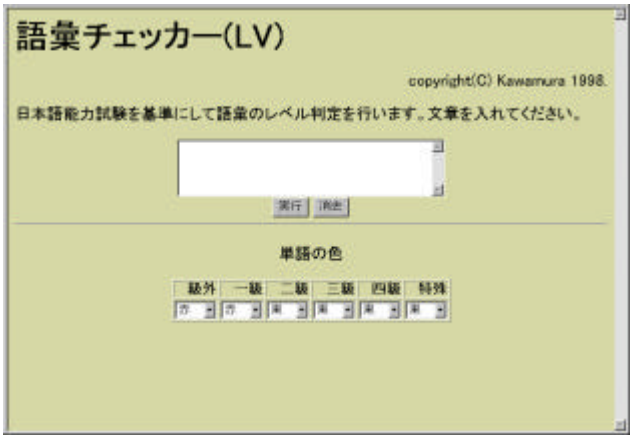

図１　語彙チェッカーの入力画面

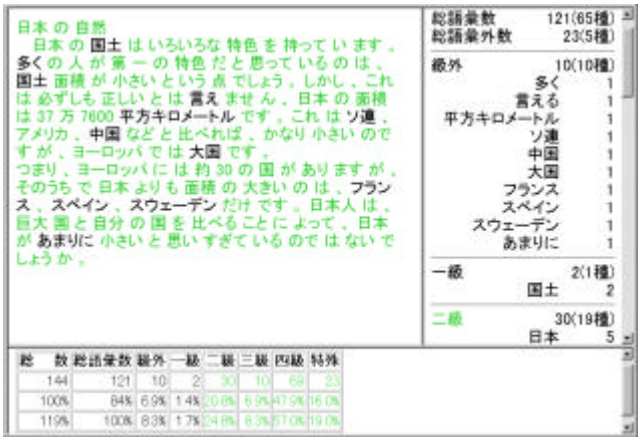

図２　語彙チェッカーの出力画面

本発表では、これらのチェッカーのうち、テキスト中の語彙の難易度を判定する「語彙チェッカー」を用いた日本語教科書の分析結果を報告する。第２節および第３節で語彙チェッカーの仕組みと活用法を概説し、第４節で語彙チェッカーを用いて行った日本語教科書の分析結果を報告する。

２．語彙チェッカーの仕組み

語彙チェッカーは、テキストに含まれるすべての語彙のレベルを日本語能力試験の１級から４級までの語彙レベルを規準に自動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。（川村　1998,1999）

図１がその入力画面である。学習者あるいは教師は、テキストを中央の枠内に打ち込むかコピー＆ペーストで張り付け、「実行」ボタンを押すだけでいい。あとはコンピュータが自動的に次の一連の作業をして、すべての語彙のレベル判定を行う。

①与えられたテキストの形態素解析を行う（形態素解析には「茶筌」を利用している）
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

図２が出力画面である。（テキストは名古屋大名古屋大学総合言語センター日本語科編『現代日本語コース中級Ｉ』第１課からとった。）画面は図２のように３種類のフレームによって構成されている。画面左上は、語彙チェッカーによる分析結果を表示したテキストである。画面右上はテキスト内の語彙を級別に分類したものであり、学習者用の語彙リストとして活用できる。画面下の表は、級別の語彙数と含有率である。実際のテキストに、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、教材の難易度と語彙の難易度との関連を調べるためのツールとして活用可能であり、本発表も、この分析結果に基づいたものである。

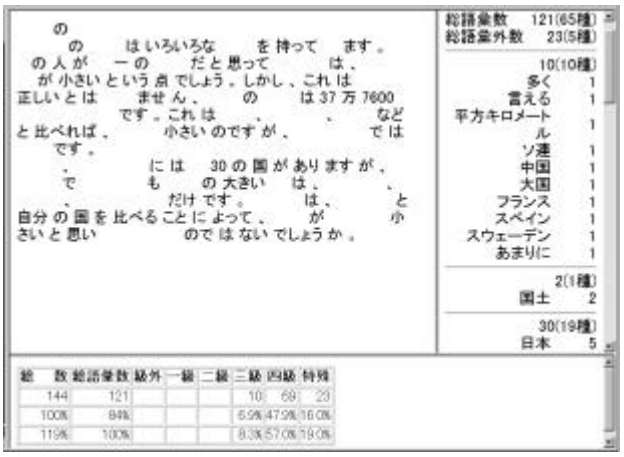

図３　２級以上の語彙を白で表示した画面

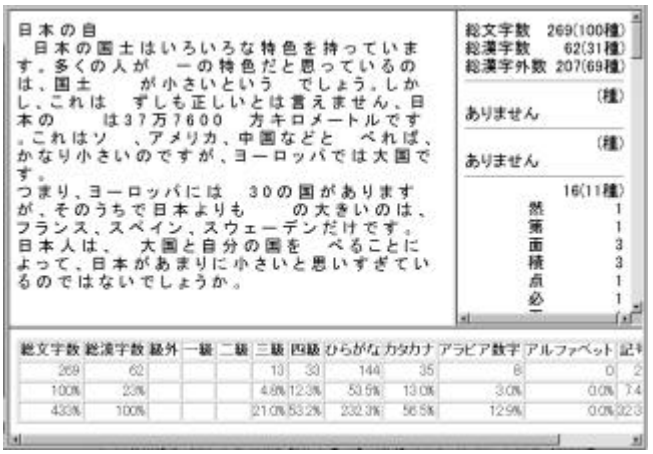

図４　２級以上の漢字を白で表示した画面

３．テキスト表示画面の活用法

では、語彙チェッカーのテキスト画面はどのような活用が可能だろうか。

まず、形態素解析の結果を知ることができる。語彙チェッカーは自動的に形態素解析を行っている。テキスト画面には、その結果が分かち書きの形で表示されているので、これを見ることで語彙チェッカーが文章を正しく分析できたかどうか確認できる。また単語毎の区切りがあるため初級学習者用テキストとして用いることも可能である。

次に、日本語能力試験の出題基準に準拠した語彙のレベル判定の結果を一目で見ることができる。語彙チェッカーの標準設定は上級学習者用（２級レベルの学習が終了）にしてある。そこで、テキスト画面では、１級以上の語彙が赤（図２では太字）で表示される。色表示はレベルごとに自由に設定できるようになっているため、どのレベルの語彙がどの程度含まれいるのか、適宜、色の設定を変えて調べることもできる。

また、この画面を活用して、学習者の「わからなさ」を疑似体験することも可能である。例えば、中級学習者を想定して、２級以上の語彙の色設定を「白」にして表示してみよう。図３がその結果表示画面である。２級以上の語彙が「白」つまり、画面では見えない状態になっている。３級レベルまでの語彙の知識のみで教材を読もうとした場合、学習者はこれと類似の状況におかれることになる。現実には、未習の語彙であっても既習の語彙からの類推で漠然と意味をつかむことができる場合もあるが、そうした方略をとらない、あるいは、とれない学習者、特に非漢字圏の学習者のおかれている状況を知ることができる。

図４は同一の教材を、漢字のレベル判定システム「漢字チェッカー」によって分析したものである。語彙チェッカーの場合と同様、２級以上の漢字を「白」で表示している。この二つの分析結果画面を比較してみると、同じ教材を用いたにもかかわらず、「わからなさ」の具合が大きく異なっていることがわかる。２級以上の漢字を抜いても、本文の内容はかなり理解できるが、２級以上の語彙をすべて消してしまうと、本文の内容は皆目見当がつかない状態になる。日本語能力試験の出題基準に準拠して読解教材を選ぶ場合、漢字を基準にするか､語彙を基準にするかでこのように大きな違いが生じてしまうことになる。

文章の難易度と語彙の難易度との関係をさらに詳しく調べるために、日本語教育の現場で用いられている日本語教科書を語彙チェッカーで分析することにした。

４．語彙チェッカーによる教科書の分析

４．１　調査対象

教科書は初級・中級・上級の教科書の中から代表的なものを数種選び、各々に含まれる語彙について語彙チェッカーを用いて判定した。対象とする教科書は、１）日本で出版されたもの、２）日本の大学への進学希望者を対象とした日本語教育の現場で用いられているもの、３）読解教材のあるものを基準に次の教科書を選定した。

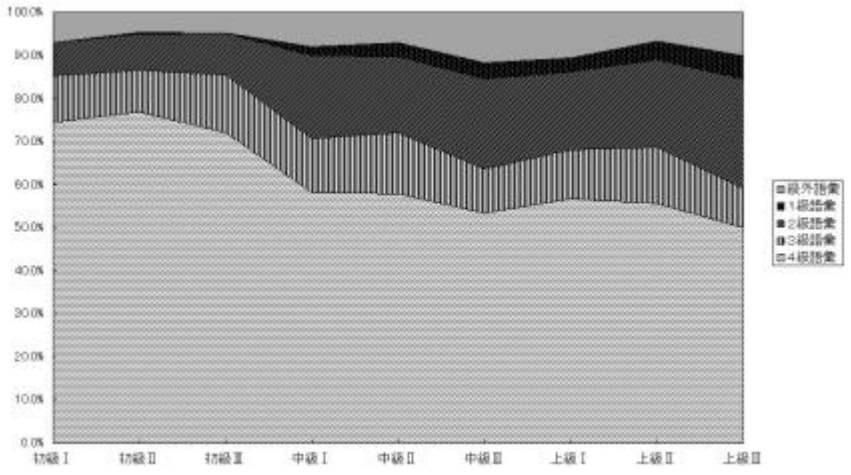

図５　テキスト全体に占めるレベル別語彙含有量（延べ語数）

初級　国際交流基金日本語国際センター『日本語初歩』
文化外国語専門学校編『文化初級日本語ⅠⅡ』
海外技術者研修協会『新日本語の基礎』
名古屋大学総合言語センター日本語科編『A Coursein Modern Japanese』

中級　国際交流基金日本語国際センター『日本語中級』
東京外国語大学留学生日本語教育センター『中級日本語』
名古屋大学総合言語センター日本語科編『現代日本語コース中級ⅠⅡ』

上級　東京外国語大学付属日本語学校編『日本語Ⅲ』

データとしては、初級に関しては本文、中級・上級に関しては読解教材部分を対象とし、級ごとに最初・中央・最後の部分を選んだ。（以下、各々を級ごとにⅠ・Ⅱ・Ⅲと表す。）対象となる文の長さが大きく異なるため、初級は３課、中級は２課、上級は１課分を分析することにした。各教材を OCR によって電子情報化したうえで語彙チェッカーにかけ、それぞれの教材に含まれる語彙のレベル判定を行った。

４．２　調査結果

調査対象となった各級の文章は初級 7,311 字、中級 14,067 字、上級 16,749 字である。形態素解析の結果、総語数は、初級 3,068、中級 6,547 語、上級 7,776 語であった。

分析の結果は図５の通りである。ただし、語数は述べ語数で算出し、レベル別の含有率を示している。この図から以下のような傾向を読みとることができる。

①４級レベルの語彙の占める割合は学習段階が進むにつれて減少傾向にある。
②３級レベルの語彙の含有率は初級から上級まで、ほぼ 10%と一定である。
③２級レベルの語彙の含有率は学習段階が進むにつれて増加する。
④１級レベルの語彙は初級ではほとんど含まれず、学習段階が進むにつれ増加傾向にある。
⑤いずれの級においても、級外の語彙が多少含まれている。初級では相対的に少ないもの

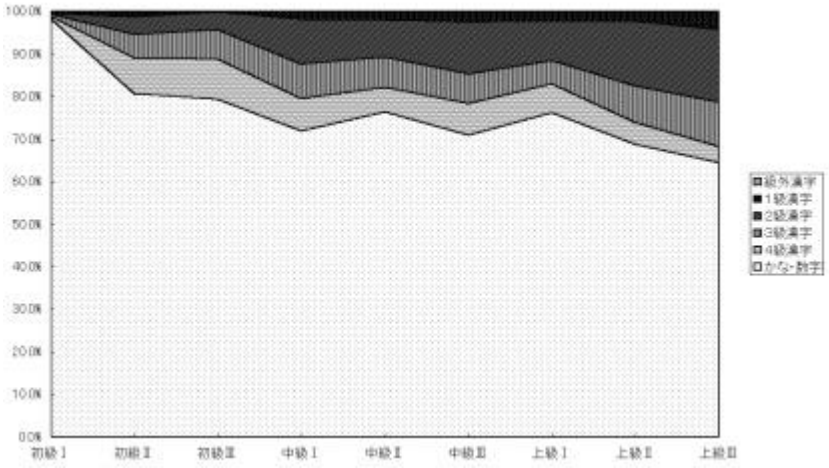

図６　テキスト全体に占めるレベル別漢字含有率（延べ文字数）

の 5%前後は含まれているし、中・上級ではいずれも平均 10%程度含まれている。

各レベルの語彙の含有率の変化に関してのピアソンの相関係数も４級は r =－ 0.74（Pearson correlation coefficient r）（N=34）と、はっきりした負の相関を示し、１級および２級はそれぞれ、 r=0.86、r=0.81 と強い正の相関を示していた。一方、級外語彙に関しては、相関係数は r=0.21 と低い。テキスト本文にあたって詳しく調べてみると、初級段階で現れる級外語彙の大半は固有名詞であった。日本語能力試験の語彙リストにはごくわずかしか固有名詞ははいっていないため、これらは級外語彙と分類されるが、学習者にとっては読み方さえわかれば理解は容易である。また、教材の内容によっては（例えば自己紹介等）固有名詞の使用頻度は高くなる可能性も大きい。語彙の難易度を決定する際、固有名詞の難易度をどう扱うかは、今後語彙チェッカーの解決すべき課題の一つである。

では、各段階の学習者にとって各々の教材はどの程度むずかしいのだろうか。例えば、中級の学習者の場合、２級以上の語彙を未習語とみなすことができる。その基準でみると、中級教科書には未習語が 30%以上含まれていることになる。また上級の学習者にとっても未習語（この場合１級以上の語彙）が 10%前後含まれている。図３で見た学習者の「わからなさ」は、以上のような未習語の高い含有率に起因していたと考えられる。

４．３　漢字チェッカーによる分析結果との比較

図６は、同じ教材を漢字チェッカーを用いて分析した結果である。総文字に対する漢字の含有率そのものは初級では少なく（1.7%～ 20.6%）、学習段階が進むにつれて、中級では 23.5%～ 29.1%、上級では 23.7 %～ 35.6%と、徐々に増えている。また、漢字全体に占める４級レベルの漢字の割合は、学習段階が進むにつれて減少傾向にあり、逆に１・２級レベルの漢字は増加傾向にある。この変化は、4.2 節でみた学習段階と語彙のレベル別含有率との関係に似た傾向を示している。

ところが、級外の語彙と漢字がテキストに果たす役割は大きく異なっている。図６が示すように、級外の語彙は初級レベルでも 5%程度含まれ、中級・上級では 10%程度含まれていた。それに対して、級外の漢字は上級の教科書であってもほとんど含まれていない。これは１級レベルの漢字リストには常用漢字がほぼ網羅されていることと、いずれの教科書も常用漢字の枠内で書かれていることによるものである。１級レベルの漢字のテキスト全体に占める割合もきわめて低い。上級の教科書であっても１級以上の漢字の含有率は数%にすぎない。また、中級の教科書も漢字の側面から分析すると中級学習者にとっての未

習漢字（２級レベル以上の漢字）は 10%前後である。つまり、語彙と漢字のいずれに着目するかによって、教材の難易度も異なってくるわけである。

では、学習者の「わからなさ」はどちらの状態なのだろうか。既習の単語や漢字知識を駆使して新しい言葉の意味を類推する能力が極めて高い学習者であれば、図４のように、未習の漢字のみがわからないことになる。当該の未習漢字の意味自体も文脈から判断することも可能であろう。しかし、単語の意味を一つ一つ確認し、覚えていくタイプの学習者にとっては、同じ文も、図３のように、見えてしまう。意味不明の単語が多すぎれば前後関係から意味を類推するどころではない。特に非漢字圏の学習者の中にこうした状態に陥っている学習者がいる可能性も否定できない。今回の調査結果は、単語の意味の類推能力を高める漢字教育の必要性を示唆するとともに、特に中級教材の作成や選択にあたって、学習者の語彙力への十分な配慮が必要なことを示していると言えよう。

５．おわりに

今回の調査に使用した語彙チェッカーは、日本語能力試験に準拠している。文章の難易度との連関をとらえる際に、語彙のレベル判定の基準として何を用いればいいかは大きな課題である。また、単語の切り出しを形態素解析に頼っているため、分析の結果に解析の精度が影響する。今後も日本語教育のために、より役立つツールとして改良を加えていく予定であるが、暫定版の語彙チェッカーをインターネット上で公開している。多くの方々に利用していただき、ご意見・ご助言をもとに、学習・教育の双方に利用しやすいツールを作り上げていきたい。

(現在、語彙チェッカーは次のURLで公開している：http://language.tiu.ac.jp)

参考・引用文献

越智洋司・矢野米雄ほか(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」『教育工学関連学協会連合第 5 回全国大会講演論文集』213-214 ページ.

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.2,10-11 ページ.

川村よし子（1999）「語彙チェッカーを用いた読解テキストの分析」『講座日本語教育』 Vol.34, 1-22 ページ.

川村よし子(1999)「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 73-87 ページ.

北村達也・川村よし子ほか(1998)「日本語読解支援システム CGI-DL における学習履歴の活用」『日本教育工学会研究報告集』35-40 ページ.

寺朱美・北村達也ほか(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.1,10-11 ページ.

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

松本裕治・北内啓ほか(1997)「日本語形態素解析システム『茶筌』ver.1.0　使用説明書,NAIST

出題基準をもとにして

	kyuugai	1kyu	2kyu	3kyu	4kyu
kanji	1%	5%	21%	30%	43%
	0%	7%	40%	27%	26%
	0%	9%	45%	26%	20%
goi	6%	0%	8%	11%	75%
	10%	3%	20%	12%	56%
	9%	4%	20%	12%	55%

	kyuugai	1kyu	2kyu	3kyu	4kyu
kanji		0.005%	0.027%	0.183%	0.540%
		0.007%	0.053%	0.162%	0.323%
		0.009%	0.060%	0.159%	0.245%
goi		0.000%	0.002%	0.013%	0.085%
		0.001%	0.004%	0.014%	0.063%
		0.001%	0.004%	0.014%	0.062%